Deleterious Background Selection With Recombination

Richard R. Hudson* and Norman L. Kaplan†

**Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717 and †Statistics and Biomathematics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709*

Manuscript received February 25, 1995
Accepted for publication August 31, 1995

ABSTRACT

An analytic expression for the expected nucleotide diversity is obtained for a neutral locus in a region with deleterious mutation and recombination. Our analytic results are used to predict levels of variation for the entire third chromosome of *Drosophila melanogaster*. The predictions are consistent with the low levels of variation that have been observed at loci near the centromeres of the third chromosome of *D. melanogaster*. However, the low levels of variation observed near the tips of this chromosome are not predicted using currently available estimates of the deleterious mutation rate and of selection coefficients. If considerably smaller selection coefficients are assumed, the low observed levels of variation at the tips of the third chromosome are consistent with the background selection model.

RECENTLY, it has been shown that the continual production of deleterious mutations along with their continual elimination by natural selection can theoretically reduce the levels of neutral variation maintained at linked loci (CHARLESWORTH *et al.* 1993). This reduction of neutral variation due to linkage to deleterious mutations is referred to as the "background selection" effect. Using Monte Carlo simulations, it was shown that the background selection effect was most pronounced in regions of low recombination (CHARLESWORTH *et al.* 1993). This raised the interesting possibility that background selection on deleterious mutations might account, at least in part, for the observation that regions of the Drosophila genome with low rates of recombination exhibit low levels of variation at the DNA level (AGUADÉ *et al.* 1989; STEPHAN and LANGLEY 1989; BEGUN and AQUADRO 1991; BERRY *et al.* 1991; BEGUN and AQUADRO 1992; MARTÍN-CAMPOS *et al.* 1992; STEPHAN and MITCHELL 1992; LANGLEY *et al.* 1993; AGUADÉ and LANGLEY 1994; AQUADRO *et al.* 1994). The theoretical analysis of CHARLESWORTH *et al.* (1993) did not give the precise dependence on recombination rates of this background selection effect. This has limited to some extent our ability to explore the possibility that background selection is playing an important role in determining the levels of variation in various parts of the Drosophila genome. Motivated by this problem, we have obtained approximate analytic results concerning the combined effects of background deleterious mutations and recombination on levels of linked neutral variation. These results, as well as simulation results to assess the accuracy of the approximations, are presented here. In addition, we illustrate how to use these analytic results by calculating the expected levels of variation as a function of physical position for the third chromosomes of *Drosophila melanogaster*, and we compare these calculated values to observed levels of variation on the third chromosome.

Corresponding author: Richard R. Hudson, Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92717. E-mail: rhudson@uci.edu

THEORETICAL ANALYSIS

Our goal in this section is to obtain an analytic expression for the mean nucleotide diversity at a neutral locus, denoted locus A, embedded in a large region in which deleterious mutations occur. Figure 1A depicts the situation being considered. Approximate results will be presented for a general model in which deleterious mutation rates and recombination rates can vary throughout the region. The positions of sites in this region are described in terms of their physical distance from locus A. For concreteness, we suppose for the moment that physical distance is measured in kb. A site x kb to the right of locus A will be said to be at position x. A site x kb to the left of locus A will be said to be at position $-x$. We denote the diploid deleterious mutation rate per generation per kb at position x by $u(x)$ and the recombination rate per generation between locus A and a site at position x by $R(x)$. We assume that recombination rates are small enough that recombination rates are additive over contiguous intervals, so that we can write $R(x) = \int_0^x r(y)\,dy$, where $r(y)$ is the local recombination rate. (That is, $r(y)\,\Delta y$ is approximately the probability of recombination per generation between the site at y and the site at $y + \Delta y$.) The far left end of the region is at position L_1, and the far right end is at position L_2. It follows that, U, the total diploid deleterious mutation rate in the region from position L_1 to L_2 equals

$$U = \int_{L_1}^{L_2} u(x)\,dx. \qquad (1)$$

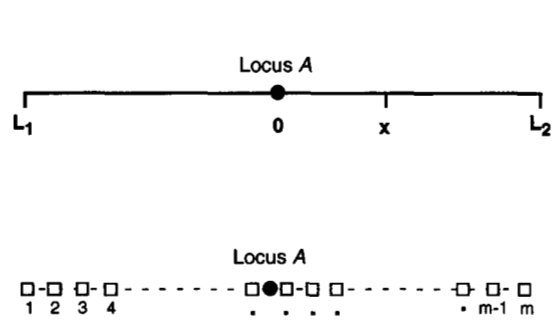

FIGURE 1.—Models considered. (A) The continuous model considered in THEORETICAL ANALYSIS. Locus A, where neutral variation occurs, is embedded in a region in which deleterious mutations occur and whose left end is at position L_1 and whose right end is at position L_2. Positions are measured in units of physical distance from locus A, with positions to the left of locus A having negative values. (That is, $L_1 < 0$ and $L_2 > 0$.) The deleterious mutation rate per kb, at a site x kb to the right of locus A is denoted $u(x)$. (B) The m-locus model that approximates the continuous model in A.

We assume that every deleterious mutation has the same selective effect, sh, and that deleterious effects combine multiplicatively. That is, an individual heterozygous for i deleterious mutations will be assumed to have fitness $(1 - sh)^i$. We assume that the selection coefficient, sh, is sufficiently large that individual mutations never reach high frequency. With these assumptions, in a very large population at equilibrium, the frequency of chromosomes with i deleterious mutations, denoted $f_i(U/2sh)$, is approximately

$$f_i(U/2sh) = \frac{(U/2sh)^i}{i!} e^{-U/2sh} \tag{2}$$

(KIMURA and MARUYAMA 1966). In other words, at equilibrium the number of deleterious mutations in the region on a random chromosome has a Poisson distribution with mean $U/2sh$. CHARLESWORTH *et al.* (1993) have shown, for an infinite-sites model with no recombination [*i.e.*, with $R(x)$ equal to zero for all $x \in (L_1, L_2)$], that the expected nucleotide diversity at locus A, denoted π, is equal to $\pi_0 \exp(-U/2sh)$, where π_0 is the nucleotide diversity that would be expected at locus A under the neutral model, in the absence of any background selection effect. [At equilibrium under an infinite-sites Wright-Fisher neutral model, π_0 is $4N\mu$, where N is the diploid population size and M is the neutral mutation rate (KIMURA 1969). We will refer to the Wright-Fisher neutral model without background selection as the strict neutral model.] Our goal is to calculate π for cases in which $R(x)$ is not zero.

We approximate the model just described by an m-locus model, as shown in Figure 1B. Properties of the desired model will be obtained by taking the limit as m tends to infinity. In the m-locus model, we number the loci from 1 to m, from left to right in the region. The ith locus has deleterious mutation rate $u(x_i)\Delta x$, where $\Delta x = (L_2 - L_1)/m$ and $x_i = i\Delta x + L_1$. It is assumed that no recombination occurs within each locus, and that the recombination rate between the neutral locus (locus A) and the ith locus is $R(x_i)$. To begin, we focus on the effects on π of deleterious mutations at a single one of the m loci, say the jth locus. That is, we suppose, for a moment, that $u(x_i)$ is zero for all values of i, except $i = j$. In this case, the problem is reduced to a two-locus model, such as that analyzed by HUDSON and KAPLAN (1994). If $u(x_j)\Delta x$ and $R(x_j)$ are small, it has been shown for the two-locus model that $\pi = \pi_0 F_j$, where

$$F_j \approx 1 - \frac{u(x_j)\Delta x \cdot sh}{2(sh + R(x_j))^2} \tag{3}$$

[from Equation 14 of HUDSON and KAPLAN(1994)]. Equation 3, which expresses the effect of deleterious mutations at a single locus on nucleotide diversity at locus A, is rederived in APPENDIX A. The problem is to determine the simultaneous affect of many such linked loci at varying distances from locus A. Remarkably, if the deleterious mutation rates at each locus are small and the population size is large, the combined effects of many loci can be obtained by simply multiplying together the effects of the individual loci. That is, π is given by

$$\pi \approx \pi_0 \prod_{i=1}^{m} F_i. \tag{4}$$

Because of this property, results are easily obtained. A derivation of (4) is given in APPENDIX A.

From (3) and (4), it follows, for the m-locus model, that

$$\pi \approx \pi_0 \prod_{i=1}^{m} \left[1 - \frac{u(x_i)\Delta x \cdot sh}{2(sh + R(x_i))^2} \right]$$

$$\approx \pi_0 \exp\left[- \sum_{i=1}^{m} \frac{u(x_i)\Delta x \cdot sh}{2(sh + R(x_i))^2} \right], \tag{5}$$

which, as m tends to infinity, approaches

$$\pi \approx \pi_0 \exp\left[- \int_{L_1}^{L_2} \frac{u(x)\, sh dx}{2(sh + R(x))^2} \right], \tag{6}$$

which is our main result. MAGNUS NORDBORG, BRIAN CHARLESWORTH and DEBORAH CHARLESWORTH (unpublished results) have obtained essentially the same result by a different approach.

We now consider some special cases. Suppose that $u(x)$ is constant and equal to u. Also, suppose that the recombination rate per kb is constant in the region. In addition, we assume that the recombination rate is sufficiently low that $R(x)$ is linear, *i.e.*, that $R(x) = r|x|$, where r is the recombination rate per kb in the region. In this case, (6) becomes

$$\pi \approx \pi_0 \exp\left[- \int_{L_1}^{L_2} \frac{ushdx}{2(sh + r|x|))^2} \right]$$

$$= \pi_0 \exp\left[- \frac{u}{2r} \cdot \left\{ \frac{rL_2}{sh + rL_2} + \frac{r|L_1|}{sh + r|L_1|} \right\} \right]. \quad (7)$$

If the neutral locus is precisely in the center of the region, then (7) simplifies to

$$\pi \approx \pi_0 \exp\left[- \frac{U}{2sh + R} \right], \quad (8)$$

where U is the total diploid deleterious mutation in the region, and R is the recombination rate between the ends of the region.

If rL_2 and $r|L_1|$ are both large compared to sh, then the term in curly brackets of (7) is approximately 2, and

$$\pi \approx \pi_0 \exp\left[- \frac{u}{r} \right], \quad (9)$$

which is the result suggested by HUDSON and KAPLAN (1994). This is also the same reduction factor that BARTON (1995) found for the effect of background selection on the probability of fixation of a weakly selected favorable mutation.

Equation 9 shows that, under certain conditions, the decrease in heterozygosity due to background selection depends on u/r but is independent of the selection coefficient against the deleterious mutations. It is important to note that this lack of dependence on selection coefficients depends on the recombination rate per unit physical distance being constant for a large region on each side of the neutral locus. For loci in regions of very low recombination this is unlikely to be true. Typically for a locus in such a region, the recombination rate increases dramatically as one moves a small map distance away from the locus (or in some cases the tip of the chromosome is encountered). Equation 9 will overestimate the background selection effect on such a locus.

To verify that (8) is a good approximation for the background selection effect on a locus in the middle of a region of uniform recombination, Monte Carlo simulations were carried out and are reported in the next section.

SIMULATIONS

Simulations very similar to those of CHARLESWORTH *et al.* (1993) were carried out to check the approximate results derived in the previous section. In these simulations, N diploid individuals were represented in the computer. Haploid genomes were represented as lists of sites at which deleterious mutations were present. In all the simulations, sites were labeled from 1 to 10,000, from left to right. To produce a descendent generation, the following procedure was carried out N times. (1) Two random individuals from the parent generation were chosen. (2) One possibly recombinant and mutated gamete was produced from each of these individuals. Each new gamete was produced without any crossovers with probability $1 - R$ and was produced via a single crossover (at a randomly chosen site) with probability R. A Poisson distributed number of new mutations were added to the gamete, at sites randomly chosen from the integers 1 to 10,000. The mean number of new mutations added per gamete was $U/2$. (3) The fitness (w) of the zygote, formed by the union of the two gametes produced in step 2, was calculated as $(1 - sh)^i$, where i is the sum of the number of deleterious mutations on the two uniting gametes. (4) With probability w the zygote was added to the growing list of the parents for the next generation, otherwise the new zygote was discarded and the program returned to step 1.

To determine the effects of background selection of deleterious mutations on linked neutral variation, a neutral locus was assumed to be located between site 5000 and site 5001 of the genomes described above. To estimate the expected nucleotide diversity at the neutral locus, a slight modification of the method of CHARLESWORTH *et al.* (1993) was used. Namely, after an equilibration period of 1000 generations, the following cycle was repeated many times. Neutral variation was introduced at the neutral locus, and evolution of the population was allowed to proceed until all variation at the neutral locus was lost. Each generation of this period, during which neutral variation persisted, the heterozygosity at the neutral locus was calculated. These heterozygosities for each generation of the cycle were summed. When all variation at the neutral locus was lost due to drift, an additional 100 generations were simulated, and then a new cycle was begun by reintroducing neutral variation. Many such cycles were carried out. To be more explicit, let us denote the heterozygosity in the tth generation of the ith cycle by z_t. We then calculated, for the ith cycle, the sum, $H_i = \Sigma\, z_t$, where the sum is over all generations of the ith cycle. If a total of M cycles were carried out, the mean and variance of the H's were estimated by

$$\hat{H} = \sum_i H_i/M \quad (10)$$

and

$$\hat{S}^2 = \sum_i^M (H_i - \hat{H})^2/M. \quad (11)$$

As described so far, the method is precisely that of CHARLESWORTH *et al.* (1993) and justification for the procedure can be found in that reference. However, in our method neutral variation was introduced at the beginning of each cycle in a way that was different from the method of CHARLESWORTH *et al.* They introduced

variation at the beginning of each cycle by introducing a single mutant at frequency $1/2N$. (They actually had 25 different neutral sites, at each one of which they introduced a single mutant allele.) With this method, the expectation of $\hat{H}$, calculated with (10), is proportional to the expected nucleotide diversity and is equal to 2 under a strict neutral model without background selection (KIMURA 1969, 1971; CHARLESWORTH *et al.* 1993). Instead of introducing a single neutral mutant, we introduce variation at the beginning of each cycle by introducing $2N$ different neutral alleles at a single neutral site, each allele having initially a frequency of $1/2N$. In our simulations, the heterozygosity (h_t) in the tth generation is calculated as $1 - \Sigma_i(p_i^2(t))$, where $p_i(t)$ is the frequency of the ith neutral allele in generation t. If variation is introduced in this way, the sum of heterozygosities, when divided by N, has the same expectation as the sum calculated by CHARLESWORTH *et al.* We found that introducing $2N$ distinct neutral alleles at the beginning of each cycle, instead of introducing a single mutant allele, reduced the number of cycles required to get good estimates of the mean summed heterozygosities. Thus, the mean summed heterozygosities calculated in this way divided by $2N$ is an unbiased estimate of the multiplicative factor by which heterozygosity is reduced by background selection. This is explicitly shown in APPENDIX B. Summarizing, we estimate π/π_0 by

$$\left\langle \frac{\pi}{\pi_0} \right\rangle = \hat{H}/2N, \tag{12}$$

and the SE of this estimate was calculated as

$$\text{SE} = \frac{\hat{S}}{2NM^{1/2}}. \tag{13}$$

Table 1 shows some simulation results along with the predicted result using (8). Overall, the agreement is excellent. With small population size ($N = 1600$), there are two cases where the simulation mean heterozygosity in the simulations is slightly, but significantly, larger than predicted mean heterozygosity. For larger population sizes, all the simulation results are within two standard errors of the predicted value, with the exception of one case (with $N = 12{,}800$) in which (8) predicts an 87% reduction in variation, whereas the simulation shows a 90% reduction in variation. We conclude that the approximation is excellent for a wide range of parameter values.

TABLE 1

Expected nucleotide diversity: a comparison of approximate theoretical expectations and simulation results

R	U	sh	N	Expected[a]	Simulation[b]
0.00	0.08	0.02	1600	0.135	0.166 ± 0.007
			3200		0.144 ± 0.007
			6400		0.143 ± 0.007
0.04	0.08	0.02	1600	0.368	0.369 ± 0.02
			3200		0.373 ± 0.02
			6400		0.357 ± 0.02
0.06	0.08	0.02	1600	0.449	0.447 ± 0.02
			3200		0.442 ± 0.02
			6400		0.425 ± 0.02
0.08	0.08	0.02	1600	0.513	0.496 ± 0.02
			3200		0.533 ± 0.02
			6400		0.528 ± 0.03
0.12	0.08	0.02	1600	0.607	0.600 ± 0.02
			3200		0.568 ± 0.02
			6400		0.638 ± 0.03
0.16	0.08	0.02	1600	0.670	0.674 ± 0.03
			3200		0.636 ± 0.03
			6400		0.726 ± 0.035
0.26	0.08	0.02	1600	0.766	0.786 ± 0.03
			3200		0.738 ± 0.03
			6400		0.786 ± 0.04
0.16	0.08	0.005	1600	0.625	0.635 ± 0.03
0.16	0.08	0.005	3200		0.635 ± 0.03
0.16	0.08	0.01	1600	0.641	0.646 ± 0.03
0.16	0.08	0.01	3200		0.639 ± 0.03
0.16	0.08	0.03	1600	0.695	0.702 ± 0.03
0.06	0.16	0.02	3200	0.202	0.192 ± 0.008
0.06	0.16	0.02	6400	0.202	0.197 ± 0.009
0.08	0.16	0.02	1600	0.264	0.260 ± 0.01
0.08	0.16	0.02	6400		0.264 ± 0.014
0.12	0.24	0.02	1600	0.223	0.243 ± 0.01
			3200		0.230 ± 0.01
			6400		0.230 ± 0.01
0.12	0.32	0.02	1600	0.135	0.167 ± 0.007
			3200		0.139 ± 0.008
			6400		0.127 ± 0.006
			12800		0.105 ± 0.004
0.16	0.32	0.02	1600	0.202	0.223 ± 0.01
0.12	0.16	0.02	1600	0.368	0.373 ± 0.016
0.16	0.16	0.02		0.449	0.457 ± 0.02

[a] Expected values are calculated with Equation 8 with $\pi_o = 1.0$.

[b] Simulation results are the means ± SE obtained from 100 cycles as described in the text.

AN APPLICATION

In this section, we illustrate the use of our analytic results by predicting levels of variation on the third chromosome of *D. melanogaster* and comparing the predictions to observed levels of variation. We will consider predictions based on (9) and also based on (6). Equation 9 requires that the recombination rate be constant for all sites within a large distance of the locus where variation is to be predicted. The more complicated (6) does not require constant recombination rates and is expected to provide more accurate predictions. Observed levels of variation, which we compare to the predictions based on (9) and (6), are estimates of $\theta(=4N\mu)$ obtained using Watterson's estimator based on the number of segregating sites (Watterson 1975) or, in some cases, estimates of nucleotide diversity (π). Both measures are appropriate to use in these comparisons since unbiased estimators of both π and θ have the same expectation under neutral models and have ap-

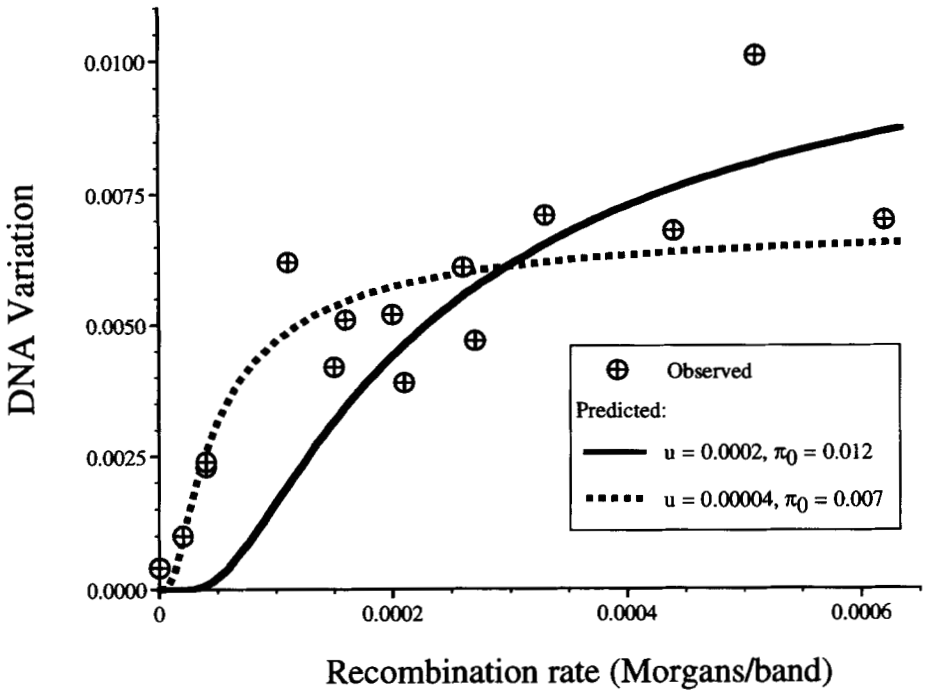

FIGURE 2.—Observed and predicted levels of DNA variation as a function of local recombination rates. The observed levels of DNA variation are estimates of θ for 15 loci on the third chromosome of *D. melanogaster* obtained from E. KINDAHL and C. AQUADRO (personal communication). These loci are, from left to right on the figure, *Lsp1γ, Pc, Antp, Gld, Ubx, tra, fz, Mlc2, ry, Sod, Tl, Rh3, Est6, E(spl),* and *Hsp26.* The local recombination rates for these loci were also provided by E. KINDAHL and C. AQUADRO. The predicted levels of variation are obtained with (9) with the parameter values indicated on the figure.

proximately the same expectation under the background selection models. [This is true, since, as argued in APPENDIX A and also HUDSON and KAPLAN (1994), the frequency spectrum of neutral variants are approximately the same under a strict neutral model and under the background selection model.] Since the estimator of θ has a lower variance than the estimator of π (NEI 1987), we have used θ when published estimates were available, and π otherwise.

Recently, E. KINDAHL and C. AQUADRO (personal communication) obtained estimates of levels of DNA variation for 15 loci on the third chromosome of *D. melanogaster.* They also estimated the local recombination rate for each of these loci using published cytological and genetic map data and plotted the observed levels of variation as a function of recombination rate. Their data are replotted in Figure 2, which shows the very strong positive relationship between level of variation and local rate of recombination noted by KINDAHL and AQUADRO. The expected dependence of level of variation on local recombination rate can be obtained with (9) if values can be assigned to the two parameters, u and π_0. The parameter u can be estimated as follows. The total diploid deleterious mutation rate in *D. melanogaster* has been estimated from mutation accumulation studies to be 1.0 or larger (CROW and SIMMONS 1983; KEIGHTLEY 1994). Since there are ~5000 cytological bands in the Drosophila genome, we estimate u, the deleterious mutation rate per generation per band, to be $1.0/5000 = 2 \times 10^{-4}$. We have no *a priori* estimate of π_0 and hence we choose its value to produce a good fit to the observed levels of variation. In Figure 2, (9) is plotted as a function of r, with $u = 2 \times 10^{-4}$ and $\pi_0 = 0.012$. The curve gives a fairly good fit to the loci with moderate to high recombination rates but predicts much lower levels of variation than are actually observed in regions of low recombination. Equation 9 is also plotted in Figure 2 with $u = 4 \times 10^{-5}$ and $\pi_0 = 0.007$, which fits moderately well over the full range of recombination rates. The parameter values, $u = 4 \times 10^{-5}$ and $\pi_0 = 0.007$, were obtained by fitting (by least squares) a straight line to the logarithm of the observed levels of variation plotted as a function of $1/r$, where r is the local recombination rate. Although a good fit with (9) is possible with $u = 4 \times 10^{-5}$, this value of u is too low to be compatible with estimated rates of deleterious mutation in Drosophila. As mentioned earlier, (9) is likely to overestimate the background selection effect for loci in regions of very low recombination. For this reason, a prediction based on (6) and using information on the entire genetic map, rather than just the local rates may be more accurate. We now calculate the expected level of variation using (6).

In *D. melanogaster* we know the approximate physical position and genetic map position of a large number of loci. With such information and a few simplifying assumptions, we can obtain an estimate of $R(x)$, the function describing the relationship between recombination distance and physical distance, that appears in (6). It is then possible to calculate π for any site in the genome using (6).

Consider the relationship between map position and physical position (expressed as number of bands from the tip of 3L) shown in Figure 3. The relationship shown in this figure is obtained from the approximate map positions of the 235 cytological subdivisions of the third chromosome of *D. melanogaster.* These map positions were taken from the file "cytotable.txt" in FLYBASE

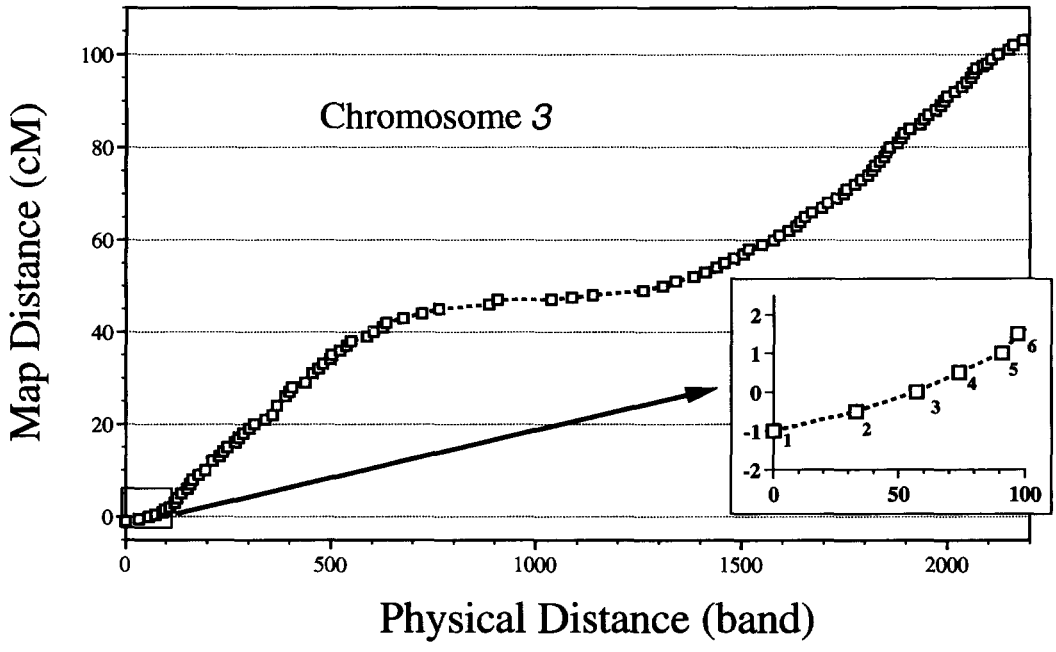

FIGURE 3.—Map distance as a function of physical position on the third chromosome of *D. melanogaster*. Physical position is measured as total number of cytological bands from the left tip of the chromosome. Map distance data are from The Drosophila Genetic Database (FLYBASE 1994). Numbers of bands for each cytological division were obtained from SORSA (1988).

(1994). This file contains an ordered list of the subdivisions with a map position associated with each subdivision. This data on map positions is rather crude and hence the predictions we obtained from them should be regarded as tentative. This map information on 235 subdivisions was reduced further by eliminating from the list all subdivisions that were indicated as having the same exact map position as the previous subdivision. This was done to avoid having intervals with estimated recombination rates of zero as an artifact of having map positions rounded to the nearest cM or, in some cases, to the nearest tenth of a cM. This procedure yielded a list of 107 subdivisions each with a map position greater than the map position of the previously listed subdivision. We associate with each of these subdivisions a single point (or locus) with physical distance from the tip of 3L given by the sum of the number of bands in all the subdivisions to the left of the subdivision. We number these loci from 1 to 107 and denote the physical position of locus i by x_i and the map position (in morgans) by $M(x_i)$. The physical and genetic position of these 107 loci are plotted as squares in Figure 3. The dashed lines connecting the squares in the figure are linear interpolations of the map positions of sites located between the loci. The interpolations represented by these dashed lines can be used to obtain an approximate $R(x)$, the function that appears in (6), for any site on the third chromosome. In this case $R(x)$ is piecewise linear and the integral in (6) is easy to calculate, if we assume that the mutation rate per band is constant throughout the chromosome. Note that we have assumed that recombination rates between loci are equal to the difference between the (approximate) map positions of the loci. This is a good approximation for tightly linked loci but not for loosely linked loci. However, the errors in our predictions due to this approximation should be small since it is only quite tightly linked loci that produce a background selection effect at a given locus.

Using (6) with $R(x)$ obtained as just described and assuming that the mutation rate per band is constant, we find that the predicted ratio, π/π_0, for locus k is given by

$$\pi/\pi_0 = e^{-G}, \tag{14}$$

where

$$G = \sum_i \frac{ush}{2} \frac{|x_{i+1} - x_i|}{(sh + |M(x_{i+1}) - M(x_k)|/2)(sh + |M(x_i) - M(x_k)|/2)}. \tag{15}$$

The terms in the denominator, $|M(x_i) - M(x_k)|$ and $|M(x_{i+1}) - M(x_k)|$, are each multiplied by $^1/_2$ in this expression to account for the fact that recombination does not occur in males. For the X chromosome one would multiply these terms by $^2/_3$, instead of $^1/_2$. Terms in this sum for which both $|M(x_i) - M(x_k)|/2$ and $|M(x_{i+1}) - M(x_k)|/2$ are much greater than sh contribute little to the sum and can be ignored. If sh is sufficiently small, then only two terms will contribute significantly to the sum, those being the terms derived from the two loci closest to the locus of interest. In this case, the result is equivalent to (9), with r replaced by an estimate of the local recombination rate obtained from the nearest mapped loci available.

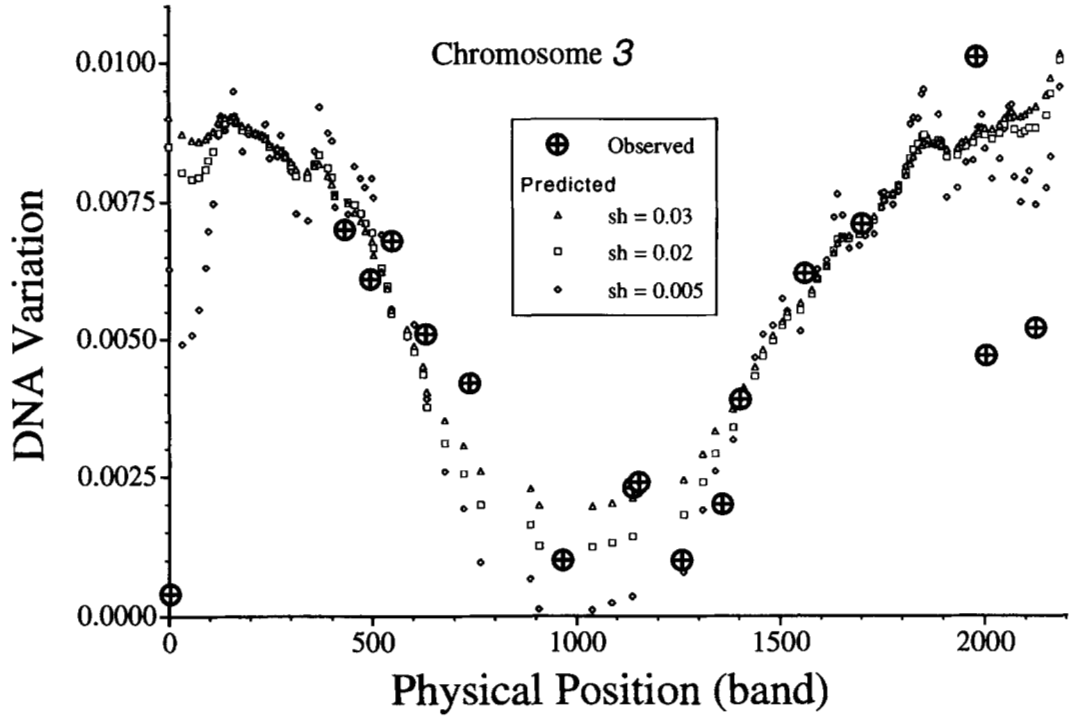

FIGURE 4.—Observed and predicted levels of DNA variation as a function of physical position on the third chromosome of *D. melanogaster*. The observed data, from left to right, are from the following loci: *Lsp1-γ, Hsp26, Sod, Est6, fz, tra, Pc, Antp,Gld, MtnA, Hsp70A, ry, Ubx, Rh3, E(spl), Tl,* and *Mlc2*. The DNA variation at *MtnA* is an estimate of π from LANGE *et al.*(1990) as reported in BEGUN and AQUADRO (1992). The observed level of DNA variation at *Hsp70A* is an estimate of π from LEIGH BROWN (1983) as reported in BEGUN and AQUADRO (1992). The other observed values are estimates of θ provided by E. KINDAHL and C. AQUADRO (personal communication). Estimates of π or θ are appropriate in this figure. Since θ has a lower variance than π, we used θ when a published estimate was available and π otherwise. The predicted values are based on (14) and (15) and assume that $\pi_0 = 0.014$ and u, the deleterious mutation rate per band, is 0.0002. The map distances shown in Figure 3 are assumed.

In Figure 4 the levels of variation observed at 17 loci on the third chromosome are plotted, in this case, as a function of physical map location rather than as a function of local recombination rates as was done in Figure 2. The prediction based on (14) is also plotted in this figure with $u = 2 \times 10^{-4}$, $\pi_0 = 0.014$ and three different values of *sh*, namely, 0.03, 0.02 and 0.005. The mutation rate of 2×10^{-4} corresponds to the previously mentioned total diploid mutation rate of 1.0 estimated from mutation accumulation studies. The *sh* value of 0.02 is the estimated average strength of selection against spontaneously occurring mutants obtained from the same studies (CROW and SIMMONS 1983). [However, see KEIGHTLEY (1994) for a more detailed analysis that allows for a distribution of selective effects.] The value of π_0, namely 0.014, was chosen to produce a good fit to the data as judged by eye. Note that with these parameter values, even loci in high recombination regions are expected to have levels of variation reduced by ~33% due to background selection. In regions of low recombination near the centromere, the reduction in variation is expected to be much greater, around a 90% reduction. This is a somewhat larger effect than calculated by CHARLESWORTH *et al.* (1993). With $\pi_0 = 0.014$, the fit of the data to the model is remarkably good except at the tips of the chromosome. There is a particularly large discrepancy at the very tip of 3L, where the observed nucleotide diversity is well below the prediction based on (6). This discrepancy is not due to approximations in the analysis of the model, which correctly incorporates the edge effects on loci near the telomere. Smaller selection coefficients would lead to greater reductions in neutral variation near the telomeres and centromeres, however, even with *sh* as small as 0.005, the predicted level of variation at the tip of 3L is well above what is observed. If $sh = 0.002$, then the predicted level of variation at the tip of 3L is close to the observed level. This value of *sh* is an order of magnitude smaller than estimated by CROW and SIMMONS (1983) and would predict lower levels of variation in the centromeric region than are actually observed.

Note also that for much of the chromosome the predicted nucleotide diversity is not strongly dependent on *sh*, consistent with (9), but in regions of low recombination, smaller *sh* leads to larger reductions in nucleotide diversity. Thus the extremely large background selection effect predicted by (9) for loci in regions of low recombination is not predicted by (6) when $sh = 0.02$ or even 0.005. Note in particular for *Lsp1-γ*, (9) predicts very low levels of variation, below what is observed, while (6) predicts relatively high levels of variation, above what is observed. This suggests that unless one believes

that $sh < 0.005$, (9) should not be used to predict levels of variation for loci such as *Lsp1-γ*, near the tips of the chromosomes or loci such as *Pc* and *Antp* near the centromere.

To predict the levels of variation on the *X* chromosome, several modifications of (14) and (15) are required. First, as mentioned earlier, the two terms in the denominator that are divided by 2 for autosomes are instead divided by 1.5 for the *X* chromosome, to account for the fact that recombination occurs only in males and the fact that *X* chromosomes occur in females two-thirds of the time instead of one-half the time for autosomes. In addition, since *X* chromosomes occur as single copies in males, the selection on mutations in the hemizygous state must be considered. If the selective effect of a deleterious mutant in hemizygous state is *s*, then *sh* must be replaced in (15) by $2sh/3 + s/3$ (CHARLESWORTH *et al.* 1993). In addition π_0 is expected to be only three-quarters as large for an *X*-linked locus as for an autosome, since in a population of *N* diploids there are only $1.5N$ copies of an *X*-linked gene, as opposed to $2N$ copies of an autosomal gene.

Our model incorporates several simplifying, but unrealistic, assumptions that should be noted, including the assumptions that all mutations have the same selective effect, that deleterious mutation rates are the same across the entire third chromosome, and that all deleterious mutations interact multiplicatively. The poor fit of the model to the data from some chromosomal regions may be due to one or more of these unrealistic assumptions, or the poor fit may be the result of other kinds of selection (*e.g.*, directional selection on advantageous mutants) acting via hitchhiking to reduce neutral variation. These issues are discussed further in the next section.

DISCUSSION

If deleterious mutations interact multiplicatively, we have shown that the effects of background deleterious selection at many loci on neutral variation at a linked locus can be calculated by simply multiplying the effects of the individual deleterious loci. This is expressed in (4) that is derived in APPENDIX A. This result leads immediately to (6) with which one can calculate the expected level of variation at a neutral locus embedded in a region with continuous recombination and deleterious mutation. To derive these results, we have considered the coalescent process that describes the genealogical history of sampled genes. Briefly, we first consider a single sampled allele at a neutral locus, which is linked to a locus at which deleterious mutations occur. The ancestor of the sampled neutral allele, *t* generations back in time, can be characterized by the number of deleterious mutations that are carried on its chromosome at the linked deleterious locus. For large *t* the distribution of the number of deleterious mutations at the linked locus is a geometric mixture of Poisson distributions, which depends on the deleterious mutation rate, selection coefficient and the recombination rate between the neutral locus and the deleterious locus. For many linked loci, each with small deleterious mutation rates, the numbers of deleterious mutations on the ancestral chromosome at the linked loci are approximately independent Poisson variables with means for each locus that depend on the recombination rate between the neutral locus and the deleterious locus. In this case, the coalescent process is straightforward. In large populations, the overall effect of the background selection is to reduce the effective population size but to otherwise preserve the neutral genealogical process. A consequence is that levels of variation are reduced, however, the frequency spectrum of variation is precisely that expected under standard neutral models. Remarkably, under some conditions the background selection effect is independent of the strength of selection against the deleterious mutations as shown in (9).

Equation 14, which is a special case of (6), can be used to predict the effect of background selection on levels of neutral variation when a detailed genetic and cytological map is available, as well as estimates of *u*, the deleterious mutation rate per band, and *sh*, the selection against individual deleterious mutations. Mutation accumulation experiments have suggested that *u* is $\sim 2 \times 10^{-4}$ per cytological band per generation, or perhaps larger, and that *sh* is ~ 0.02 or perhaps somewhat smaller in *D. melanogaster* (CROW and SIMMONS 1983). With these parameter values we have shown that the background selection effect is expected to be substantial on all loci of the third chromosome. Even at loci in regions of high recombination on this chromosome, the predicted level of variation is approximately two-thirds the level expected without background selection. Near the centromere of the third chromosome, the background selection model predicts substantial reductions in levels of variation. Available data are consistent with this prediction. However, using current estimates of mutation rates and selection coefficients, predictions for the tips of the third chromosome are not consistent with available data. Loci near the tips of the third chromosome exhibit lower levels of variation than are predicted under the background selection model. Variation on the fourth chromosome is also lower than expected under the background selection model (CHARLESWORTH *et al.* 1993).

It is important to consider the robustness of our quantitative predictions to violations of a number of simplifying, but unrealistic, assumptions upon which (14) is based. For example, we assume an infinite-sites model with constant population size, a complete lack of interference of recombination, multiplicative interaction of deleterious mutations, and that all deleterious mutations have the same effect. We now discuss these in turn.

Under the infinite-sites model, multiple hits, by assumption, do not occur. In reality, there are a finite

number of sites at any locus and each site has at most four possible states and hence multiple hits are possible. For some organisms and some loci, multiple hits will be common either due to high mutation rates or very long coalescent times, and in these cases a finite-sites model, possibly with variation of mutation rates across sites, will need to be considered. Extending our results to these more realistic finite-sites models will be straightforward given our results concerning the coalescent times. In fact, since we have shown that the distribution of genealogies with background selection is approximately the same as under strict neutral models, the results for any finite-site neutral models would apply to the corresponding background selection model with the appropriate reduction in effective population size. We note, however, that in *D. melanogaster* there is little evidence of multiple hits in samples taken from within the species, and hence we believe that an infinite-site model is appropriate for interpreting polymorphism data in *D. melanogaster*.

In our theoretical analysis, we have assumed a completely linear recombinational map without interference. This is not likely to introduce much error into our predictions, since the loci more than a few map units apart are unlikely to have an effect on each other under the background selection model. Over these small distances nonlinearities and interference will have very minor effects. It should be noted that our simulations are based on an assumption of complete interference. BRIAN CHARLESWORTH (personal communication) is currently carrying out simulations with a more realistic model of recombination that should alleviate many concerns about the recombination assumptions used here.

All of our analysis is based on the assumption that fitness is multiplicative across loci. Given the possibility that synergistic interactions may be common (KONDRASHOV 1988; CHARLESWORTH 1990), it is important to regard our results as tentative in this regard. CHARLESWORTH *et al.* (1993) suggested that synergism may increase the effect of background selection. It is highly desirable that models with synergism and other forms of interaction be investigated further.

Our predictions using (14) are based on a rather crude genetic map, but these predictions are unlikely to change due to small refinements in the map, unless selection coefficients are smaller than the values that we have assumed. However, if selection coefficients are much smaller, then much greater effects of background selection are expected in regions of low recombination. Also, in this case, the effects are expected to be quite sensitive to small changes in the genetic map. For example, if the map position of the tip of *3*L is assumed to be -0.5 instead of -1.0, as we have assumed, and if $sh = 0.002$, then the predicted nucleotide diversity at the tip of the chromosome is 0.0012. This predicted level of variation is certainly not significantly different from the observed level of variation at *Lsp1-γ*, near the tip of *3*L. [E. KINDAHL and C. AQUADRO (personal communication) estimate θ for this locus to be 0.0004.] Thus to account for the observed patterns of variation at the tips of the chromosomes, it appears that higher deleterious mutation rates or smaller selection coefficients are required. Smaller selection coefficients might result in the very large background selection effects observed at the tips of the chromosomes and on chromosome *4*. However, such low selection coefficients would result in levels of variation lower than are observed near the centromere of the third chromosome at least with the crude map data that we have used. If selection coefficients are typically as low as 0.002, then much more refined genetic maps may be necessary to predict accurately the expected level of variation under the background selection model.

Another possibility is that deleterious mutation rates and selection coefficients vary on a large scale across the chromosomes. For example, transposable element insertions are plausibly slightly deleterious mutations (GOLDING 1987), and they are nonrandomly distributed in the genome (CHARLESWORTH *et al.* 1992). Hence, as suggested earlier by HUDSON (1994), transposable elements may result by themselves in heterogeneous deleterious mutation rates and thus produce a background selection effect that is nonuniform across the genome.

In addition, we have assumed that all deleterious mutations have exactly the same deleterious effect, *sh*. It is more reasonable to assume that there is a distribution of deleterious effects. Generalizing (6) to the case where *sh* has a distribution is straightforward. When recombination rates are zero, the harmonic mean of *sh* can be substituted for *sh*, as pointed out earlier (CHARLESWORTH *et al.* 1993), but when recombination rates are not zero, such a substitution is not appropriate. BRIAN CHARLESWORTH (personal communication) is currently carrying out a detailed analysis for the third, second and *X* chromosomes, using more precise genetic maps and including the effects of transposable elements and distributions of selective effects. The predictions based on this more detailed analysis should be much more precise, especially in regions of low recombination.

In conclusion, though parameter values of u and sh are uncertain, and the nature of the interaction between deleterious mutations is uncertain, it seems very likely that background selection is having an important effect on levels of variation in some regions of the *D. melanogaster* genome. Background selection can account for the very low levels of variation observed at loci near the centromere of the third chromosome. However, the very low levels of variation observed near the tips of the third chromosome and on the fourth chromosome may require additional explanations, although the background selection model with higher deleterious mutation rates or lower selection coefficients in these regions could account for the observations. We emphasize that the frequency spectrum of

neutral variants under the background selection model is approximately the same as under a strict neutral model. This is in contrast to the hitchhiking of advantageous mutants model that has also been proposed as an explanation of the low levels of variation observed in regions with low recombination (KAPLAN *et al.* 1989; WIEHE and STEPHAN 1993). Under the hitchhiking of advantageous mutants model, low frequency variants are expected to be more common than under a neutral model (BRAVERMAN *et al.* 1995). A global analysis of the frequency spectrum of variants in different regions of the *D. melanogaster* genome may be quite informative for distinguishing between the background selection and the hitchhiking models.

Many properties of genetic variation depend on the effective population size. The analysis that we have carried out suggests that effective population size may be very different for different regions of the Drosophila genome. This will influence equilibrium levels of variation as described above, but many other properties as well. For example, we expect measures of geographic structure, such as F_{st}, for loci in regions of low recombination could be much larger than for loci in regions of high recombination because the product of effective population size and migration rate will be much smaller in regions of low recombination. Similarly, aspects of the evolution of a population that depends on the product of population size and selection coefficient are likely to be different in regions low recombination than in regions of high recombination. Thus background selection may dramatically effect the evolutionary process in large portions of the genome.

This work was supported in part by U. S. Public Health Service grant GM-42397. We also thank BRIAN and DEBORAH CHARLESWORTH for helpful discussions and for suggesting that predictions for entire chromosomes would be useful. MAGNUS NORDBORG, BRIAN CHARLESWORTH and DEBORAH CHARLESWORTH also generously shared unpublished results.

LITERATURE CITED

AGUADÉ, M., and C. H. LANGLEY, 1994 Polymorphism and divergence in regions of low recombination in Drosophila, pp. 67–76 in *Non-neutral Evolution: Theories and Molecular Data,* edited by G. B. GOLDING. Chapman & Hall, New York.

AGUADÉ, M., N. MIYASHITA and C. H. LANGLEY, 1989 Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster.* Genetics **122:** 607–615.

AQUADRO, C. F., D. J. BEGUN and E. C. KINDAHL, 1994 Selection, recombination and levels of DNA polymorphism in Drosophila, pp. 46–56 in *Non-neutral Evolution: Theories and Molecular Data,* edited by G. B. GOLDING. Chapman & Hall, New York.

BARTON, N. H., 1995 Linkage and the limits to natural selection. Genetics **140:** 821–841.

BEGUN, D. J., and C. F. AQUADRO, 1991 Molecular population genetics of the distal portion of the X chromosome in Drosophila: evidence for genetic hitchhiking of the yellow-achaete region. Genetics **129:** 1147–1158.

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

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

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

CHARLESWORTH, B., 1990 Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet. Res. **55:** 199–221.

CHARLESWORTH, B., A. LAPID and D. CANADA, 1992 The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster.* I. Element frequencies and distribution. Genet. Res. **60:** 103–114.

CHARLESWORTH, C., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

CROW, J. F., and M. J. SIMMONS, 1983 The mutation load in Drosophila, pp. 1–35 in *The Genetics and Biology of Drosophila,* edited by M. ASHBURNER, H. L. CARSON and J. N. THOMPSON. Academic Press, London.

FLYBASE, 1994 The Drosophila Genetic Database. Available from the ftp.bio.indiana.edu network server and Gopher site.

GOLDING, G. B., 1987 The detection of deleterious selection using ancestors inferred from a phylogenetic history. Genet. Res. Camb **49:** 71–82.

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

HUDSON, R. R., and N. L. KAPLAN, 1994 Gene trees with background selection, pp. 140–153 in *Non-neutral Evolution: Theories and Molecular Data,* edited by G. B. GOLDING. Chapman & Hall, New York.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

KEIGHTLEY, P. D., 1994 The distribution of mutation effects on viability in *Drosophila melanogaster.* Genetics **138:** 1315–1322.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

KIMURA, M., 1971 Theoretical foundations of population genetics at the molecular level. Theor. Popul. Biol. **2:** 174–208.

KIMURA, M., and T. MARUYAMA, 1966 The mutational load with epistatic gene interactions in fitness. Genetics **54:** 1337–1351.

KONDRASHOV, A. S., 1988 Deleterious mutations and the evolution of sexual reproduction. Nature **336:** 435–440.

LANGE, B. W., C. H. LANGLEY and W. H. STEPHAN, 1990 Molecular evolution of Drosophila metallothionein genes. Genetics **126:** 921–932.

LANGLEY, C. H., J. MACDONALD, N. MIYASHITA and M. AGUADÉ, 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans.* Proc. Natl. Acad. Sci. USA **90:** 1800–1803.

LEIGH BROWN, A. J., 1983 Variation at the *87A* heat shock locus in *Drosophila melanogaster.* Proc. Natl. Acad. Sci. USA **80:** 5350–5354.

MARTÍN-CAMPOS, J. M., J. P. COMERON, N. MIYASHITA and M. AGUADÉ, 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster.* Genetics **130:** 805–816.

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

SORSA, V., 1988 *Chromosome Maps of Drosophila.* CRC Press, Inc., Boca Raton, FL.

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

STEPHAN, W., and S. J. MITCHELL, 1992 Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae.* Genetics **132:** 1039–1045.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **10:** 256–276.

WIEHE, T. H. E., and W. STEPHAN, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster.* Mol. Biol. Evol. **10:** 842–54.

Communicating editor: A. G. CLARK

APPENDIX A

To make the paper self-contained we rederive (3) in a way that is more appropriate for the problem in this paper. We then derive (4). Let T denote the time (measured in generations) until two randomly chosen chromosomes at the neutral locus A have a common ancestor. If μ is the neutral mutation rate, then under an infinite-sites model

$$\pi = 2\mu E(T) = \pi_0 \frac{E(T)}{2N}, \tag{A1}$$

where N is the diploid population size and $\pi_0 = 4N\mu$. HUDSON and KAPLAN (1994) calculated $E(T)$, assuming the neutral locus A was linked to a single deleterious locus, using a set of recurrence equations. We now describe an alternative approach that can be generalized to more than one linked deleterious locus.

Following the notation in the text, we define for $i = 1, 2$

$X_i(t)$ = the number of deleterious mutations at locus j on the ancestral chromosome in the tth ancestral generation, of the ith sampled chromosome.

X_1 and X_2 have the same distribution, and if the population size is very large, then we can also assume that they are independent. Since all chromosomes having k deleterious mutations are selectively equivalent, the probability, Λ_t, that the two chromosomes have a common ancestor in generation t equals

$$\Lambda_t = \sum_k \frac{P(X_1(t) = k)^2}{2Nf_k(u(x_j)\Delta x/2sh)}, \tag{A2}$$

where we recall that for $x > 0$ and $k \geq 0$,

$$f_k(x) = \frac{e^{-x}x^k}{k!}$$

and $u(x_j)\Delta x$ is the deleterious mutation rate at locus j. $2Nf_k(u(x_j)\Delta x/2sh)$ is the number of chromosomes at equilibrium with k mutations at locus j (KIMURA and MARUYAMA 1966). If N is large and time is measured in units of $2N$ generations, then it follows from (A2) that the distribution of T is approximately exponential with parameter

$$\Lambda_\infty \approx \sum_k \frac{P_\infty(k)^2}{f_k(u(x_j)\Delta x/2sh)}, \tag{A3}$$

where

$$P_\infty(k) = \lim_{t\to\infty} P(X_1(t) = k),$$

assuming the limit exists. Hence $E(T) \approx (\Lambda_\infty)^{-1}$.

The same line of reasoning applies if there is more than one deleterious locus. Indeed if there are m such loci, then

$$\Lambda_\infty \approx \sum_{\underline{k}} \frac{P_\infty(\underline{k})^2}{\prod_{j=1}^m f_{k_j}(u(x_j)\Delta x/2sh)}, \tag{A4}$$

where $\underline{k} = (k_1, k_2, \ldots, k_m)$ and $\prod_{j=1}^m f_{k_j}(u(x_j)\Delta x/2sh)$ is the frequency of chromosomes at equilibrium with k_j mutations at locus j, $j = 1, \ldots, m$. The problem thus reduces to studying the behavior of

$$P_\infty(\underline{k}) = \lim_{t\to\infty} P(X_1(t) = (k_1, k_2, \ldots, k_m)).$$

We return to the case of one deleterious locus and determine the stationary distribution of $X_1(t)$. To simplify notation we employ u for $u(x_j)\Delta x$ and R for $R(x_j)$. We assume that R is small enough so that each gamete has at most one recombination event between the two loci per generation. Each generation a randomly chosen gamete is either a recombinant or not, with probabilities R and $1 - R$, respectively, and has Y_1 deleterious mutations that were present in the parent and Y_2 new mutations. Y_1 and Y_2 are independent Poisson variables with means $u/2sh$ and $u/2$, respectively. It follows from properties of Poisson variables that the distribution of Y_1 conditional on the sum $Y_1 + Y_2$ is binomial with one parameter being $Y_1 + Y_2$, and the other parameter equal to the mean of Y_1 divided by the sum of the mean of Y_1 and Y_2. Thus conditional on $X_1(t)$, $X_1(t + 1)$ has with probability $1 - R$, a binomial distribution with parameters $X_1(t)$ and $(u/2sh)/(u/2sh + u/2) = 1/(1 + sh)$, and with probability R, a Poisson distribution with mean $u/2sh$. Let $h_t(\beta) = E(\beta^{X(t)})$, $0 < \beta < 1$. It follows from the above that $h_{t+1}(\beta)$ satisfies the recursion

$$h_{t+1}(\beta) = (1 - R)h_t(g(\beta)) + Re^{-u(1+sh)(1-\beta)/2sh}, \tag{A5}$$

where

$$g(\beta) = 1 - \left(\frac{1}{1 + sh}\right)(1 - \beta).$$

We can iterate (A5) to obtain the stationary solution as

$$h_\infty(\beta) = R\sum_{k=0}^{\infty} (1 - R)^k e^{-u(1+sh)(1-\beta)/[2sh(1+sh)^k]}. \tag{A6}$$

It follows from (A6) that the stationary distribution is a geometric mixture of Poisson distributions.

If $u/2sh$ and sh are small, then we can approximate $h_\infty(\beta)$ as

$$h_\infty(\beta) \approx R\sum_{k=0}^{\infty} (1 - R)^k\left(1 - \frac{u(1 - \beta)}{2sh(1 + sh)^k}\right)$$

$$\approx 1 - \frac{uR(1 - \beta)}{2sh(R + sh)}.$$

Hence the stationary distribution of X_1 satisfies

$$P_\infty(1) = \frac{uR}{2sh(R + sh)} \text{ and } P_\infty(0) = 1 - P_\infty(1). \tag{A7}$$

Substituting (A7) in (A3) and doing some algebra leads to

$$\Lambda_\infty \approx 1 + \frac{ush}{2(R + sh)^2}$$

and so finally we obtain (3)

$$E(T) \approx \Lambda_\infty^{-1} \approx 1 - \frac{ush}{2(R + sh)^2}. \qquad \text{(A8)}$$

We next indicate how to calculate Λ_∞ for more than one deleterious locus. For simplicity we assume that there are two such loci, but the argument generalizes to any number. Let $u_1 = u(x_1)\Delta x$ and $u_2 = u(x_2)\Delta x$, be the mutation rates at the two loci and $R_1 = R(x_1)$ and $R_1 + R_2 = R(x_2)$ be the recombination rates between each of the loci and the neutral locus. [we assume that $R(x_1) < R(x_2)$.] The generalization of (A6) is straightforward and leads to the following recursion

$$\begin{aligned} h_{t+1}(\beta_1, \beta_2) &= (1 - R_1 - R_2)h_t(g(\beta_1), g(\beta_2)) \\ &\quad + R_1 e^{-u_1(1+sh)(1-\beta_1)/2sh} e^{-u_2(1+sh)(1-\beta_2)/2sh} \\ &\quad + R_2 h_t(g(\beta_1)) e^{-u_2(1+sh)(1-\beta_2)/2sh}. \end{aligned} \qquad \text{(A9)}$$

Making the same approximations that we did to derive (A7), we obtain

$$\begin{aligned} h_\infty(\beta_1, \beta_2) &\approx 1 - \frac{u_1(1 - \beta_1)}{2sh} \\ &\quad \times \left(\frac{R_1}{R_1 + R_2 + sh} + \frac{R_1 R_2}{(R_1 + R_2 + sh)(R_1 + sh)} \right) \\ &\quad - \frac{u_2(1 - \beta_2)}{2sh} \left(\frac{R_1 + R_2}{R_1 + R_2 + sh} \right) = 1 - \frac{u_1(1 - \beta_1)}{2sh} \\ &\quad \times \left(\frac{R_1}{R_1 + sh} \right) - \frac{u_2(1 - \beta_2)}{2sh} \left(\frac{R_1 + R_2}{R_1 + R_2 + sh} \right). \end{aligned} \qquad \text{(A10)}$$

Extending (A10) to larger numbers of loci is straightforward. The only conditions that must be satisfied is that the associated recombination rates all be different and that $u(x)$ be well behaved, *e.g.*, piecewise continuous. A reasonable condition to impose on $R(x)$ is that it can be written as

$$R(x) = \int_0^x r(y)\, dy,$$

where $r(y) > 0$. If $R(x)$ is constant over a finite interval, *i.e.*, there is a recombinational cold spot, then the deleterious mutation rate need not be small and so (A7) does not necessarily hold.

It follows from (A10) that at stationarity the numbers of deleterious mutations on the ancestral chromosome in disjoint intervals $I_1, \ldots, I_m$ are nearly independent Poisson variables with means

$$\int_{I_j} \frac{u(x)R(x)}{2sh(R(x) + sh)}\, dx, \quad j = 1, \ldots, m.$$

An immediate consequence of the independence property is that

$$P_\infty(\underline{k}) = \prod_{j=1}^{m} P_\infty(k_j), \qquad \text{(A11)}$$

where $\underline{k} = (k_1, \ldots, k_m)$ and $P_\infty(k_j)$ is given by (A7) with the appropriate value of R. The proof of (4) now follows by simple algebra.

It is straightforward to generalize the analysis to larger samples. In particular, if a random sample of size n is obtained from a large population, the time back until the first coalescent event, measured in units of $2N$ generations, is approximately exponentially distributed with mean equal to

$$\left[\binom{n}{2} \Lambda_\infty \right]^{-1}.$$

Any of the $\binom{n}{2}$ possible pairs of lineages are equally likely to coalesce at the first coalescent event. These are precisely the properties of a gene genealogy of a sample of size n, under a strict neutral model, if Λ_∞ is set equal to 1 (TAJIMA 1983). Thus, the coalescent process under the background selection model is approximately the same as under a strict neutral model, except that Λ_∞ is given by (A4) under the background selection model, whereas Λ_∞ is equal to 1 under the neutral model. Summarizing, if the population is sufficiently large, the distribution of gene genealogies of a sample is approximately the same under background selection model and under the neutral model, except that the effective population size under the background selection model is reduced by the factor

$$\text{Exp}\left[- \int_{L_1}^{L_2} \frac{u(x)sh}{2(sh + R(x))^2}\, dx \right].$$

An important consequence of this result is that the sample frequency spectrum is not affected by background selection.

APPENDIX B

In this appendix, we justify using (12) for estimating the mean nucleotide diversity expected under the background selection model relative to the mean nucleotide diversity under a strict neutral model. From (A1) it follows that (12) would provide an unbiased estimate of π/π_0, if $E(\hat{H})$ is equal to $E(T)$, where T is the time (measured in generations) until two randomly chosen chromosomes have a common ancestor at a neutral locus. We now show that $E(\hat{H}) = E(T)$.

It is well known that $E(T) = \Sigma_t P(T > t)$, where $P(T$

$> t$) denotes the probability that T is greater than t. (We assume here that T is a discrete random variable.) With Monte Carlo simulations, we can estimate $P(T > t)$, for a particular value of t, as follows. Assuming that we have a population at stationarity at the loci at which deleterious mutations occur, we introduce neutral variation at the neutral locus as described in the main text, *i.e.*, $2N$ distinct alleles are introduced, each with frequency $1/2N$. We run the simulation an additional t generations with mutation and selection continuing to occur at the background loci, but without mutation at the neutral locus. Since no additional mutations have occurred at the neutral locus, two sampled alleles at the neutral locus have a common ancestor less than t generations back only if they are identical. It follows that the expectation of the heterozygosity after t generations is equal to $P(T > t)$. In the notation of SIMULATIONS, z_t is the heterozygosity in the tth generation, and hence, $E(z_t) = P(T > t)$. It follows that $H_t = \Sigma\, z_t$ (and $\hat{H}$) have means equal to $E(T)$. Notice that in the simulations, once all the variation is lost at the neutral locus, the heterozygosity is zero from then on, so the sum, H_t, does not grow anymore, and there is no reason to continue summing the z_t's. Instead one can reintroduce neutral variation and obtain another realization of the random variable $\Sigma\, z_t$.