Genetic Drift in an Infinite Population: The Pseudohitchhiking Model

John H. Gillespie

Section of Evolution and Ecology, University of California, Davis, California 95616

Manuscript received June 26, 1999
Accepted for publication February 3, 2000

ABSTRACT

Selected substitutions at one locus can induce stochastic dynamics that resemble genetic drift at a closely linked neutral locus. The pseudohitchhiking model is a one-locus model that approximates these effects and can be used to describe the major consequences of linked selection. As the changes in neutral allele frequencies when hitchhiking are rapid, diffusion theory is not appropriate for studying neutral dynamics. A stationary distribution and some results on substitution processes are presented that use the theory of continuous-time Markov processes with discontinuous sample paths. The coalescent of the pseudohitchhiking model is shown to have a random number of branches at each node, which leads to a frequency spectrum that is different from that of the equilibrium neutral model. If genetic draft, the name given to these induced stochastic effects, is a more important stochastic force than genetic drift, then a number of paradoxes that have plagued population genetics disappear.

THIS article investigates the hypothesis that linked selection rather than genetic drift is the major stochastic force in many natural populations. Certain kinds of linked selection can produce stochastic dynamics that are remarkably like those of genetic drift. If true, this hypothesis may explain a number of paradoxical observations about genetic variation in natural populations.

The ideas presented here have at least four main antecedents. The first, of course, is Maynard Smith and Haigh's (1974) seminal article on "the hitchhiking effect." Their investigation was prompted by a problem raised in Lewontin (1974): Assuming that protein variation is neutral, "the extent of enzyme polymorphism is surprisingly constant between species." So constant, in fact, that the effective sizes of most species must be within one order of magnitude of each other. Maynard Smith and Haigh argued that hitchhiking events are like population bottlenecks in their ability to reduce genetic variation to levels that will be similar across species. This article is an exploration of that idea.

The second antecedent is the extensive literature showing that genetic variation is reduced in regions of low recombination (Aguadé *et al.* 1989; Miyashita 1990; Berry *et al.* 1991; Begun and Aquadro 1992; Aguadé and Langley 1994). The simplest hypothesis to explain this phenomenon is the effects of linked selection. While there is an active controversy over the form of this selection (Kaplan *et al.* 1989; Charlesworth *et al.* 1993; Charlesworth 1994; Braverman *et al.* 1995; Gillespie 1997), there is general agreement over the hypothesis that some form of linked selection causes the reduction.

The third antecedent is a simulation study that showed that adaptive substitutions can cause the level of genetic variation at a linked neutral locus to be only weakly dependent on the population size (Gillespie 1999). This simulation confirms the basic premise in Maynard Smith and Haigh (1974) that hitchhiking can cause a homogenization of levels of variation across species, but points out that for this to happen, the rate of substitution at the selected locus must be an increasing concave function of population size.

The fourth antecedent came from Will Provine during a conversation in Liberia, Costa Rica, in which he tried to convince me that genetic drift must be a minor force compared to the effects of linked selection. He used an asexual haploid species to make his point, but his arguments carry weight for sexual, diploid species and are a major impetus for the work reported here.

The main goal of this article is to describe the effects of a steady stream of adaptive substitutions at one locus on the dynamics of a linked, neutral locus. A full mathematical treatment of this situation is out of reach. However, it appears that the induced stochastic effects of the substitutions on the neutral locus can be faithfully captured in a one-locus model called the *pseudohitchhiking model*.

NO CROSSING-OVER

We begin with the study of a neutral locus that is so tightly linked to a selected locus that there is no crossing-over between them. Both the selected and the neutral loci are represented by Watterson's infinite-sites, no-recombination model of a gene (Watterson 1975). Evolution occurs in a finite population of size N subject to the standard assumptions of the Wright-Fisher model.

The mutation rate at the neutral locus is called u and

Author e-mail: jhgillespie@ucdavis.edu

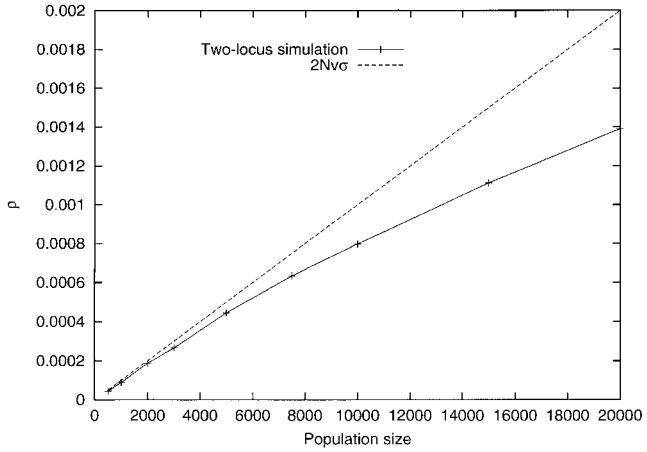

Figure 1.—The rate of substitution, ρ, at the selected locus where the selection coefficient is $\sigma = 0.1$ and the mutation rate is $v = 5.0 \times 10^{-7}$.

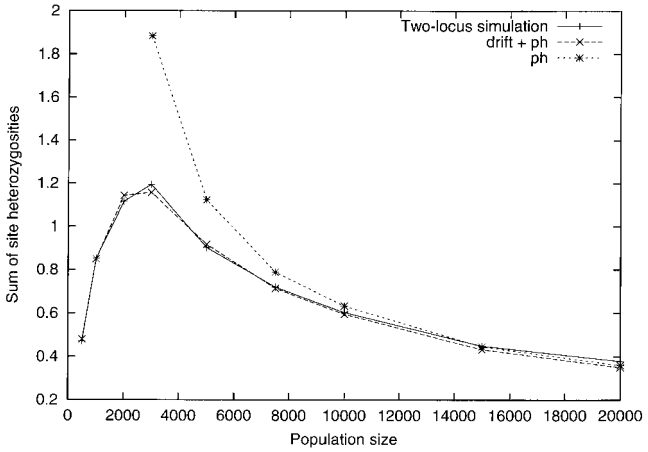

Figure 2.—The average sum of site heterozygosities, ssh, for the neutral locus linked to the selected locus illustrated in Figure 1. The neutral mutation rate is $u = 2.5 \times 10^{-4}$. The drift + ph curve comes from Equation 1 and the ph curve comes from Equation 2.

the mutation rate at the selected locus is called v. Each mutation at the selected locus raises the fitness of the homozygote for that mutation by an amount σ over the fitness of the homozygote for the parent allele. The heterozygote fitness is exactly intermediate between the two homozygote fitnesses. In accordance with the assumptions of the shift model (Ohta and Tachida 1990), all fitnesses are measured relative to that of the allele with the most recently fixed site.

The rate of fixation of advantageous mutations at the selected locus, ρ, as a function of the population size is illustrated in Figure 1. These results were obtained from a computer simulation using the same approach and lisp code as described in Gillespie (1999). The values of ρ obtained from this simulation play an important role in what follows. The usual approximation for this rate, $2Nv\sigma$, is also plotted.

Our main interest is in the properties of the neutral locus that is linked to the selected locus. Variation at the neutral locus is measured by the sum of site heterozygosities (SSH),

$$\text{SSH} = \sum_{i=1}^{S} 2x_i(1 - x_i),$$

where S is the number of segregation sites at the neutral locus and x_i is the frequency of one of the two mutations at the ith site. For an isolated neutral locus the mean value of SSH, which we call ssh, is

$$\text{ssh} = E\{\text{SSH}\} = 4Nu.$$

Figure 2 gives the average sum of site heterozygosities of the neutral locus that is linked to the selected locus described in Figure 1. After an initial rise, ssh falls slowly with increasing population size. At first this seems paradoxical because, for an isolated neutral locus, ssh increases linearly with population size. The reason for this contrary behavior is that the rate of substitution at the selected locus increases with population size (see Figure 1), which lowers the variation at the linked neutral locus by an amount that overcomes the increased mutational input.

It is a simple matter to describe the relationship between ssh and N mathematically if we make two simplifying assumptions:

1. The times of fixations at the selected locus form a Poisson process with rate ρ.
2. The time required to fix a selected allele is so short relative to the time between substitutions ($1/\rho$) and the time scale of genetic drift (N) that the fixations may be viewed as occurring instantaneously.

With these two assumptions, the mean time back to the common ancestor of a pair of randomly chosen neutral alleles is

$$\frac{1}{\rho + 1/2N} = \frac{2N}{1 + 2N\rho}.$$

The mean number of mutations on the two lineages leading to the common ancestor is just $2u$ times the mean time back to the common ancestor, or

$$\text{ssh} = \frac{4Nu}{1 + 2N\rho}. \tag{1}$$

This formula is plotted in Figure 2 and is indistinguishable from the simulated values except for the largest population sizes, where it is too low. (The reason for the lack of agreement for large N is discussed later.) As the population size increases,

$$\lim_{N\to\infty} \text{ssh} = \frac{2u}{\rho}, \tag{2}$$

which is also illustrated in Figure 2. The second formula converges rather quickly to the simulated values, an

observation that could, of course, be gleaned from Equation 1 itself. Biologically, this observation suggests that the stochastic effects of linked selection completely dominate those of genetic drift once the population size is $> \sim 10^4$. By extrapolation, in an infinite population we would still have a stochastic force affecting the neutral locus, but that force would not be genetic drift. We now want to argue that this force, although not genetic drift, shares many properties with genetic drift.

In an infinite population, three fates await a neutral allele whose frequency is x_i in a given generation:

1. A selected mutation that ultimately fixes in the population could appear on the same chromosome as one copy of our allele and that copy would then be whisked to fixation by hitchhiking. The probability that a favorable mutation appears on a copy of our allele is just its frequency, x_i.
2. A selected mutation that ultimately fixes in the population could appear on the same chromosome as some other allele. In this case our allele will be eliminated from the population.
3. No selected mutation that ultimately fixes enters the population, in which case the frequency of our allele remains unchanged.

The frequency of our allele, after a hitchhiking event that may have occurred has run its course, may be summarized as follows:

$$x_i' = \begin{cases} 1 & \text{with probability } \rho x_i \\ 0 & \text{with probability } \rho(1 - x_i) \\ x_i & \text{with probability } 1 - \rho. \end{cases}$$

The change in x_i is

$$\Delta x_i = \begin{cases} (1 - x_i) & \text{with probability } \rho x_i \\ -x_i & \text{with probability } \rho(1 - x_i) \\ 0 & \text{with probability } 1 - \rho. \end{cases}$$

The mean and variance in Δx_i are

$$E\{\Delta x_i\} = 0$$
$$\mathrm{Var}\{\Delta x_i\} = \rho x_i(1 - x_i).$$

The variance in Δx_i is of the same form as that for genetic drift. This is the first hint that the stochastic effects of linked selection share some properties with those of genetic drift.

In a finite population, the variance in the change of x_i becomes

$$\mathrm{Var}\{\Delta x_i\} = x_i(1 - x_i)\left(\frac{1}{2N} + \rho\right). \quad (3)$$

This formula immediately suggests that the stochastic effects of linked selection can be viewed as formally no different than those of genetic drift, but with a population size reduced to

$$N_e = \frac{N}{1 + 2N\rho}. \quad (4)$$

Consistent with this view is the fact that $4N_e u$ is equal to the value of ssh given in Equation 1. However, while this suggestion is accurate for some properties (like ssh), it is off the mark for others. For example, Figure 5 shows that Tajima's *D*-statistic (Tajima 1989), a measure of departure from the neutral frequency spectrum, is negative when both stochastic effects are present.

In an infinite population, the stochastic effects of linked selection on a neutral locus can be examined mathematically, but not by using diffusion theory as is often done to study genetic drift. Rather, we use a continuous-time Markov model with discontinuous sample paths. That diffusion theory is not appropriate follows from

$$E\{(\Delta x_i)^3\} = \rho x_i(1 - x_i)(1 - 2x_i),$$

which is similar in magnitude to $\mathrm{Var}\{\Delta x_i\}$. To use diffusion approximations, $E\{(\Delta x_i)^3\}$ must be of a smaller order of magnitude than $\mathrm{Var}\{\Delta x_i\}$.

Many of the sorts of problems that have been solved for genetic drift have analogs in this new context. For example, consider the stationary distribution of a neutral locus with two alleles that mutate to one another with rate u and are linked to a selected locus with substitutions occurring at rate ρ. As the only force acting between hitchhiking events is mutation, the frequency of one of the alleles is given by

$$x(\tau) = \frac{1}{2}(1 - e^{-2u\tau}) + x(0)e^{-2u\tau},$$

where τ is the time since the last hitchhiking event and $x(0) = 1$ if the allele was the one fixed at the last event and $x(0) = 0$ if it was not. Suppose that the latter happened [so $x(0) = 0$], then

$$\mathrm{Prob}\{X_t < x\} = \mathrm{Prob}\{T < -\ln(1 - 2x)/2u\},$$

where X_t is the frequency of the allele at time t and T, which is exponentially distributed with rate ρ, is the time back to the most recent hitchhiking event. Thus,

$$\mathrm{Prob}\{X_t < x\} = 1 - (1 - 2x)^{\rho/2u}.$$

From here it is a simple matter to show that the density for x is

$$\frac{\rho}{2u}|1 - 2x|^{\rho/2u - 1}.$$

This is the pseudohitchhiking analogue to the β-density that describes the balance between drift and mutation under the Wright-Fisher model.

A closely related density, which can be compared to the simulations, is that for the frequency of the unmutated copies of the most recently fixed neutral allele. After a fixation τ generations ago, the frequency of unmutated copies of the allele is

$$e^{-u\tau}.$$

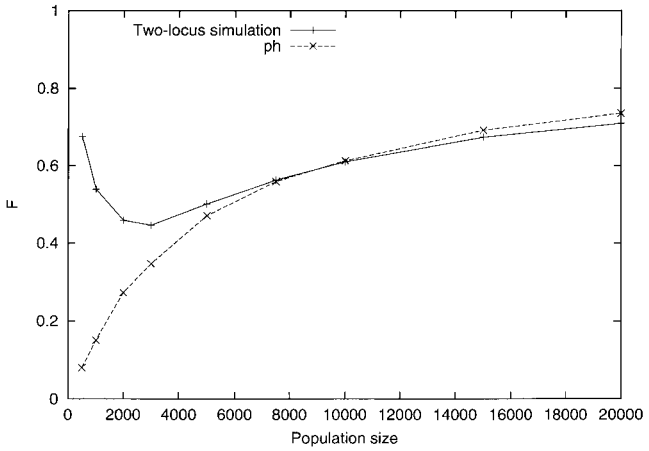

Figure 3.—The allelic homozygosity, F, for the neutral locus linked to the selected locus illustrated in Figure 1. The neutral mutation rate is $u = 2.5 \times 10^{-4}$. The ph curve comes from Equation 5.

The probability that this frequency is $<x$ is

$$\text{Prob}\{e^{-u\tau} < x\} = \text{Prob}\{T > -\ln x/u\},$$

where, as before, T is exponentially distributed with rate ρ. Thus, the density of x is

$$(\rho/u)x^{\rho/u-1}.$$

Under the assumption that all new mutations are unique, the homozygosity of the population is

$$F = \int_0^1 x^2(\rho/u)x^{\rho/u-1}dx = \frac{1}{1+2u/\rho}. \qquad (5)$$

This result is the analogue of the well-known expression for the homozygosity of a neutral population under mutation and drift,

$$\frac{1}{1+4N_e u},$$

which points out that $2u/\rho$ under the pseudohitchhiking model plays the same role as $4Nu$ under the neutral model. Equation 5 can also be derived by using the value of N_e given into Equation 4 in the previous formula and then taking the limit as $N \to \infty$.

Equation 5 is compared to the two-locus simulations in Figure 3. As the population size grows, the simulations and the equation come into close agreement. The reason that they differ for smaller population sizes is that Equation 5 assumes an infinite population size. Once the population size is large enough, the agreement between the simulations and theory is very good. At the largest population sizes, the two curves begin to diverge. The reason for this appears to be that the assumption that substitutions occur instantaneously breaks down. For example, when $v = 5 \times 10^{-7}$ and $N = 10{,}000$, the fixation time of a selected allele is 438 generations and the time between substitutions is ~1265 generations. Thus, the fixations are reasonably spaced. However, at $N = 20{,}000$, the fixation time is ~504 generations and the time between fixations is ~719 generations. In this case, there is considerable overlap between the substitutions, which violates the time-scale assumption and leads to a higher than expected homozygosity. Similarly, we saw in Figure 2 that ssh is too low when the population size is very large.

A proper limit could be obtained from the two-locus model by allowing $N \to \infty$ as $v \to 0$ in such a way that ρ remains constant. Unfortunately, we do not have an explicit formula for the dependency of ρ on N, so we cannot state the conditions required for convergence as $N \to \infty$. However, were we willing to accept the usual approximation,

$$\rho \approx 2Nv\sigma,$$

then

$$v = \frac{\rho}{2N\sigma}$$

should provide the proper scaling. By holding v fixed with increasing population size in Figure 5, we necessarily depart from the pseudohitchhiking model as $N \to \infty$.

Another problem that is easily handled concerns the fixation process for the neutral locus. Recall that the origination process is made up of the times of appearance of mutations that ultimately fix in the population and the fixation process is the times that they ultimately fix (Gillespie 1993). The origination process for the neutral model is a Poisson process with rate u, this being so even if there is linked selection. Watterson (1982, 1984) gave some partial results for the neutral fixation process, which is considerably more complicated than the origination process as multiple sites may fix in the same generation. In particular, he was able to show that the number of sites that fix in a particular generation, given that at least one site fixed, is geometrically distributed. He was not able to find the distribution of the times between these fixation episodes.

In an infinite population, the time between fixation events may be written as the sum of two random times. The first of these, Y, is the time until the first appearance of a mutation that will ultimately fix. As the origination process is Poisson, this time is exponentially distributed with rate u. The second, Z, is the time until the next hitchhiking event (which must of necessity fix the mutation). By assumption, Z is exponentially distributed with rate ρ. Thus, the times of the fixation episodes form a renewal process with the time between events, usually called the failure time, being $T = Y + Z$. The moments of T are

$$E\{T\} = \frac{u+\rho}{u\rho} \qquad (6)$$

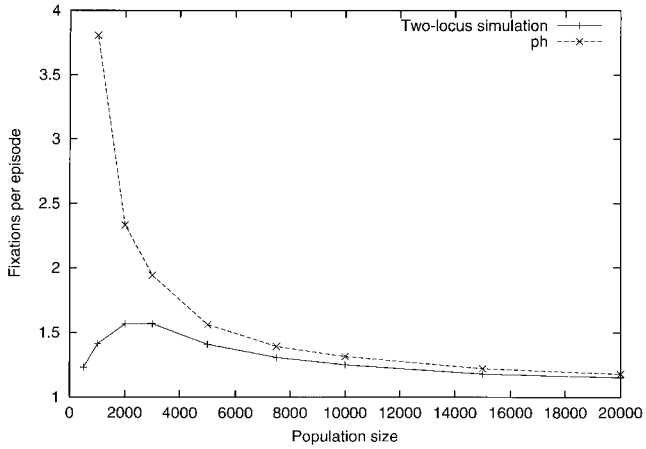

Figure 4.—The number of sites fixed per fixation episode for the neutral locus linked to the selected locus illustrated in Figure 1. The ph curve comes from Equation 8.

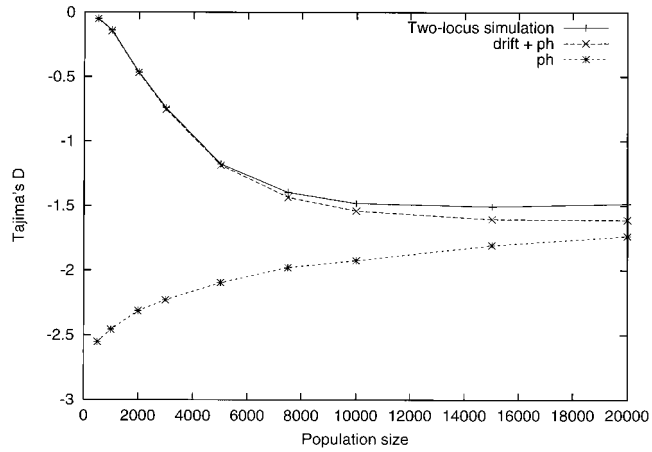

Figure 5.—Tajima's D for a sample of size 20 from the neutral locus linked to the selected locus illustrated in Figure 1. The drift + ph curve comes from a direct simulation of the coalescent for the pseudohitchhiking model as described in Equation 9 and the ph curve is the same but with $N = \infty$ for all population sizes.

$$\mathrm{Var}\{T\} = \frac{u^2 + \rho^2}{u^2\rho^2}. \tag{7}$$

The asymptotic index of dispersion for a renewal process is the variance in the failure time divided by the square of the mean failure time (Cox 1962). For our case, this is

$$R = \frac{u^2 + \rho^2}{(u + \rho)^2}.$$

For example, in the simulations for the case $N = 20{,}000$ and $u = 2.5 \times 10^{-4}$, we observed that $\rho = 1.39 \times 10^{-3}$ and $R = 0.8027$. The previous formula gives $R = 0.7969$, which, once again, is in very close agreement.

The number of mutations that fix, given that at least one fixes, is 1 plus a random number whose distribution is a Poisson randomized by uZ. As a Poisson randomized by an exponential is geometrically distributed, we have that the number of sites that fixes at each episode is geometric with mean

$$1 + u/\rho. \tag{8}$$

Figure 4 illustrates that this mean agrees very well with those observed in the simulations. Note that the mean number of mutations that fix in an interval of length t is

$$\frac{t}{E\{T\}}(1 + u/\rho) = ut,$$

as expected. It is worth noting that this is the only fully characterized fixation process known at this time.

The final observation concerns the nature of the coalescent. In a finite population, the coalescent will be the usual neutral coalescent until the first hitchhiking event, at which point all of the extant lineages coalesce. The death process for the number of extant lineages, n, is governed by the following transition probabilities:

$$n \to \begin{cases} n & \text{with probability } 1 - \rho - \dfrac{n(n-1)}{4N} \\ n-1 & \text{with probability } \dfrac{n(n-1)}{4N} \\ 1 & \text{with probability } \rho. \end{cases} \tag{9}$$

The properties of this coalescent, while amenable to a mathematical analysis, are most easily studied by a direct simulation of the coalescent using a slight generalization of the method described by Hudson (1990). Here, we also add mutations to the coalescent and calculate Tajima's D-statistic as a way of reducing the properties of the coalescent to a single number. The results of such a simulation for both finite and infinite populations are compared to those of the two-locus simulation in Figure 5. The agreement is quite good for smaller population sizes, but the two-locus simulation and the direct coalescent simulations diverge for larger population sizes. The reason for the divergence is clearly the problem of overlapping selected substitutions as discussed above.

In this section we have been comparing a two-locus simulation to calculations that flow from a pair of assumptions: the Poisson nature of the selected substitution process and the instantaneous fixation time of selected substitutions. These two assumptions allow us to model the behavior of the neutral locus without any explicit use of the dynamics of the selected locus other than knowledge of the value of ρ. We call this one-locus model the *pseudohitchhiking model*, the prefix *pseudo* serving to emphasize that the full hitchhiking dynamics are not part of the one-locus model. We have seen that the properties of the pseudohitchhiking model are very close to those of the two-locus model when the assumptions of the model are met. In particular, we require

that the hitchhiking events form an isolated stream of impulses.

CROSSING-OVER

When crossing-over occurs between the selected and neutral loci, selection no longer carries the hitchhiking allele to fixation, which requires a modification of the pseudohitchhiking model. In this section we consider a modification that gives acceptable results for tight linkage. The development of the model itself is quite instructive and provides insights into factors that must be considered in developing a more sophisticated version.

In the first step of the generalization of the pseudohitchhiking model, we allow the frequency of the neutral hitchhiking allele to stop before reaching fixation. When a favorable mutation first enters the population, it is on the same chromosome as only one copy of a neutral allele. The frequency of that copy will increase from $1/2N$ to some new value, call it y, at the expense of all other copies of the allele and all other alleles, which will have their frequencies reduced by a fraction $1 - y$. Thus, after a hitchhiking event has run its course, the frequency of the neutral alleles will have changed according to the following scheme:

$$x_i' = \begin{cases} (1-y)x_i + y & \text{with probability } \rho x_i \\ (1-y)x_i & \text{with probability } \rho(1-x_i) \\ x_i & \text{with probability } 1-\rho. \end{cases}$$

The values of x_i' do not exactly match the verbal description above because we have failed to reduce the frequency of the ith allele from x to $x - 1/2N$ to reflect the changed status of the one copy of that allele that is linked to the selected mutation. This small perturbation can be included and shown to alter the calculations below by a quantity of order $(1/2N)^2$, which is deemed negligible.

The change in x_i is

$$\Delta x_i = \begin{cases} y(1-x_i) & \text{with probability } \rho x_i \\ -yx_i & \text{with probability } \rho(1-x_i) \\ 0 & \text{with probability } 1-\rho, \end{cases}$$

and the mean and variance in Δx_i are

$$E\{\Delta x_i\} = 0$$
$$\mathrm{Var}\{\Delta x_i\} = \rho y^2 x_i(1-x_i).$$

Note that this model reduces to that of the previous section when $y = 1$. When drift is added,

$$\mathrm{Var}\{\Delta x_i\} = x_i(1-x_i)\left(\frac{1}{2N} + \rho y^2\right). \tag{10}$$

From this we see that the effective size of the population is

$$N_e = \frac{N}{1 + 2N\rho y^2}$$

and that the mean sum of site heterozygosities is

$$\mathrm{ssh} = \frac{4Nu}{1 + 2N\rho y^2}. \tag{11}$$

At this point we have reached an impasse: What is the value of y? If we choose to make y a parameter of the model, then its value can be derived from the results of Maynard Smith and Haigh (1974). But y will surely be a random variable reflecting the stochastic dynamics of hitchhiking. In this case, the values of y associated with a sequence of hitchhiking events will form a stationary sequence of independent, identically distributed random variables. Unfortunately, there is no available theory that allows us to derive the distribution of y. Before addressing the problem of adding randomness to the model, we examine the model with a deterministic y and use this as a benchmark to measure further refinements in the model.

Maynard Smith and Haigh (1974) describe the effects of the substitution of a new advantageous mutation at one locus on the frequency of two neutral alleles at a linked locus. The new mutant is originally on the same chromosome as one of the two neutral alleles and causes the frequency of that allele to increase by an amount that is determined by three parameters: the selection coefficient, σ, the rate of recombination, r, and the population size, N. (The latter parameter is relevant only through the assumption that the frequency of the newly arisen selected mutation is $1/2N$; there is no genetic drift in their model.) The final frequency of the neutral allele that was linked to the advantageous mutation is

$$x_\infty = 1 - r(1-x_0)(1-p_0)\int_0^\infty \frac{e^{-rz}dz}{1 - p_0 + p_0 e^{\sigma z}}, \tag{12}$$

where x_0 was the frequency of the hitchhiking allele in the population before the advantageous allele appeared and $p_0 = 1/2N$ is the initial frequency of the advantageous mutation. This result, as presented, is a continuous-time additive diploid version of Equation 8 in the Maynard Smith and Haigh (1974) article.

Equation 12 can be rearranged as

$$x' = x(1-y) + y, \tag{13}$$

where

$$x = x_0$$
$$x' = x_\infty$$

and

$$y = 1 - r(1-p_0)\int_0^\infty \frac{e^{-rz}dz}{1 - p_0 + p_0^{\sigma z}}, \tag{14}$$

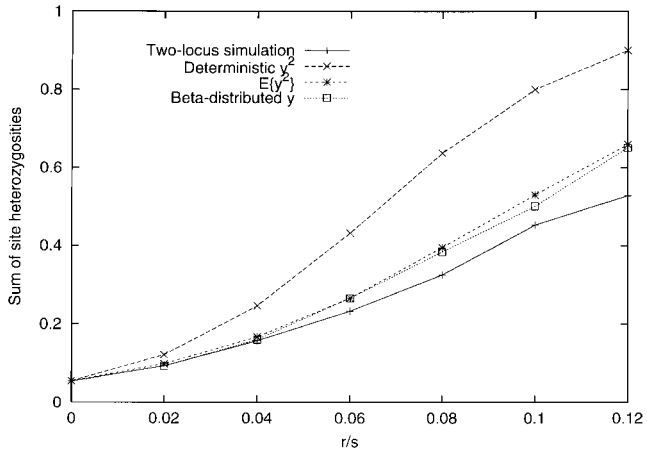

Figure 6.—Simulation results comparing the sum of site heterozygosities (ssh) for the neutral locus in the two-locus model to that of three versions of the pseudohitchhiking model with those of a neutral locus linked to a selected locus. In all three simulations $N = 20{,}000$, $4Nu = 1.0$, $\sigma = 0.1$, and $v = 1.5 \times 10^{-7}$. The pseudohitchhiking simulations use $\rho = 4.29 \times 10^{-4}$, as obtained from the two-locus simulations. The deterministic y^2 curve uses values of y obtained from Equation 14. The $E\{y^2\}$ curve uses values of $E\{y^2\}$ obtained from a simulation of the Maynard Smith and Haigh model as described in the text. Beta-distributed y curve uses a β-distribution for y.

which could provide a value for y in the pseudohitchhiking model. However, rather than obtaining y by numerical integration of Equation 14, it is easier to obtain the value from a deterministic simulation of the additive diploid version of the Maynard Smith and Haigh difference equations.

Figure 6 presents values of ssh for a two-locus simulation as well as those using Equation 11 with the value of ρ taken from the selected locus in the two-locus simulation and the value of y obtained numerically as described above. The agreement is good, as it must be, for very tight linkage. However, for weaker linkage the two-locus and pseudohitchhiking results diverge significantly. (For very loose linkage they will converge again as hitchhiking becomes unimportant.)

To add randomness, we first note that the values of y form a sequence of independent identically distributed random variables, and thus that

$$\mathrm{Var}\{\Delta x_i\} = x_i(1 - x_i)\left(\frac{1}{2N} + \rho E\{y^2\}\right) \qquad (15)$$

and

$$\mathrm{ssh} = \frac{4Nu}{1 + 2N\rho E\{y^2\}}. \qquad (16)$$

These observations suggest that randomness may require nothing new other than the substitution of $E\{y^2\}$ for y^2. The values for $E\{y^2\}$ can be obtained from a simulation of the Maynard Smith and Haigh model in a finite population using the same parameters as used in the two-locus simulation. The third curve in Figure 6 is a plot of Equation 16 with the values of $E\{y^2\}$ coming from the Maynard Smith and Haigh simulation. The agreement between this version of the pseudohitchhiking model and the two-locus simulation is much better than that using the deterministic value of y^2. The improvement comes from the fact that the $E\{y\}$ from the finite-population simulation is considerably larger than the deterministic value of y. The reason for this appears to be that in finite populations the sample path of an advantageous mutation, given that it fixes, has a mean trajectory that lies above the trajectory of the deterministic process. Conditioning on fixation will favor sample paths that move rapidly away from zero, for obvious reasons.

Of course, the full dynamics of the pseudohitchhiking model will depend on the complete distribution of y rather than on just $E\{y^2\}$. We can examine a complete model by assuming that y is β-distributed and obtaining its mean and variance from the same Maynard Smith and Haigh two-locus simulations that were described in the preceding two paragraphs. This distribution is then used in a direct simulation of the pseudohitchhiking model from which the average value of ssh is recorded. The results of these simulations are illustrated in the fourth curve in Figure 6. There is essentially no difference between these results and those from Equation 16. This is not unexpected as ssh for neutral models depends only on the first and second-order moments in the change in the neutral allele frequencies and these, in turn, depend only on the first two moments of y. Other properties of the pseudohitchhiking model may well depend on higher-order moments of the process and will require the use of a sequence of random values of y rather than the fixed value $E\{y^2\}$.

The agreement between the pseudohitchhiking model with random y and the two-locus simulations for larger values of r is still not as good as we would hope. There are two potential sources for the discrepancy. The first is that the dynamics of selected substitutions in the Maynard Smith and Haigh simulations are not identical to those of the two-locus simulations. In the latter, which is an infinite-sites model, there are always several alleles segregating at the selected locus. In fact, sometimes two advantageous alleles with the same fitness will move through the population at the same time. Such dynamics are considerably more complicated than those of the Maynard Smith and Haigh simulations, which always have exactly two segregating alleles at the selected locus. At this time I have no way to assess the impact of this difference.

A second source for the discrepancy concerns the assumption that the frequencies of all of the nonhitchhiking alleles in the pseudohitchhiking model are lowered by the same constant factor $1 - y$. When y is deterministic, this is the correct assumption. However, when y is random it is not correct to assume that all of the nonhitchhiking alleles are lowered by the same factor.

In fact, the nonhitchhiking alleles should all be lowered by a different random amount, reflecting the various effects of drift and recombination that occur during the hitchhiking event. In some cases, there may even be two separate hitchhiking alleles. It appears to be quite difficult to add this particular element of randomness, although further work may uncover a way.

Although further refinements of the pseudohitchhiking model will be forthcoming, the remainder of this article is concerned with the properties of the model as defined above.

THE COALESCENT

As a first step, consider the genealogy of n alleles sampled from a pseudohitchhiking population with deterministic y and $N = \infty$. In this case, the only way that a coalescence can occur is if there is a hitchhiking event. The probability of such an event in a particular generation is ρ. If there were an event, then a single copy of one of the alleles in the population increases its frequency to y. The probability that i of the n sampled alleles are descended from that fortunate allele is the binomial probability

$$\binom{n}{i} y^i (1 - y)^{n-i}.$$

A coalescence occurs when $i \geq 2$. Unlike the neutral case, a coalescence can involve more than two lineages, which is the root cause of $D < 0$.

We can summarize these observations as follows:

The probability that a coalescence does not occur in a particular generation is

$$(1 - \rho) + \rho[(1 - y)^n + ny(1 - y)^{n-1}].$$

The probability that a coalescence does occur in a particular generation is

$$\rho \sum_{i=2}^{n} \binom{n}{i} y^i (1 - y)^{n-i}.$$

The probability that the coalescent shrinks from n alleles to $n - i$ alleles in a particular generation is

$$\rho \binom{n}{i+1} y^{i+1} (1 - y)^{n-i-1}.$$

When $n = 2$, the probability of a coalescence is ρy^2. Thus, the mean number of mutations separating these two alleles is

$$\frac{2u}{\rho y^2}. \tag{17}$$

This same result can be obtained by taking the limit of Equation 11 as $N \to \infty$:

$$\lim_{N \to \infty} \frac{4Nu}{1 + 2N\rho y^2} = \frac{2u}{\rho y^2}.$$

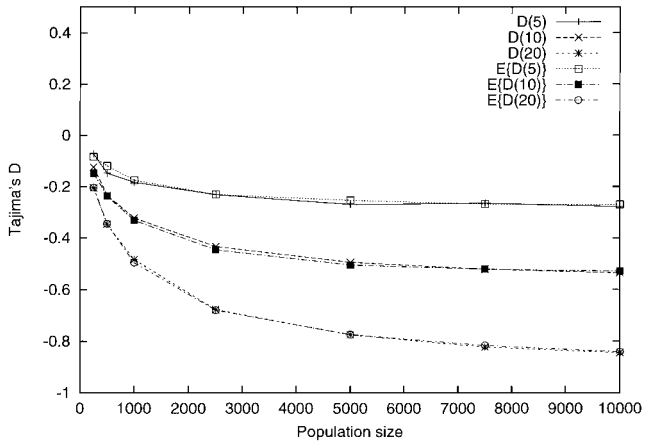

Figure 7.—The average values of Tajima's D for different sample sizes. Those $E\{D(n)\}$ curves come from samples drawn from a direct simulation of the pseudohitchhiking model for a sample of size n. The $D(n)$ curves come from a direct simulation of the coalescent using Equation 18. In both cases, $\rho = 0.138$, $y = 0.3$, and $u = 5 \times 10^{-4}$.

The next increment in complexity involves the addition of genetic drift. In any particular generation, a coalescence may be due to the finiteness of the population or to hitchhiking. In the former case, the coalescent can only shrink from n to $n - 1$ while in the latter case the size of the coalescent can shrink from n to $n - i$, $i = 1 \ldots (n - 1)$. Thus, the probabilities of all possible transitions are

$$n \to \begin{cases} n & \text{w.p. } 1 - \rho - \dfrac{n(n-1)}{4N} - \rho[(1-y)^n + ny(1-y)^{n-1}] \\ n-1 & \text{w.p. } \dfrac{n(n-1)}{4N} + \rho\binom{n}{2} y^2 (1-y)^{n-2} \\ n-i & \text{w.p. } \rho\binom{n}{i+1} y^{i+1} (1-y)^{n-i-1}, \quad i = 1 \ldots (n-1). \end{cases} \tag{18}$$

These transition probabilities, plus the usual assumption that the times to successive coalescences are exponentially distributed, allow a complete probabilistic description of the coalescent for small n. However, for $n > 4$ the results are completely unwieldy. On the other hand, it is very easy to simulate the coalescent in the same manner that is done for neutral coalescents (Hudson 1990).

Figure 7 gives examples of the calculation of Tajima's D using a direct simulation of the pseudohitchhiking model and using a coalescent simulation with the transition probabilities given above. The two approaches give identical answers, as they should. There are two interesting aspects to these results. The first is that Tajima's D becomes more negative with increasing population size. The negativity comes from the fact that a coalescence can involve more than two lineages, the increasing magnitude comes from a decreasing role of genetic drift and with it, a decreasing frequency of $n \to n - 1$ transitions.

The second interesting aspect of Figure 7 is the in-

crease in D that accompanies increasing sample sizes. Tajima's D for the entire population is -2.2 when $N = 10{,}000$. Thus, as the sample size increases, D does approach the population value. However, the approach is not sufficiently fast for D to be a reliable estimator of the population D. D has a dual role as an estimator and for hypothesis testing. D is scaled such that $|D| > 2$ is cause to reject the neutral model. However, it is clear from Figure 7 that even though a population may have a large skew in its frequency spectrum, D will not exhibit a mean value that is close to the significance level for the sort of sample sizes typically used in population studies. Thus, there is reason to doubt that D has sufficient power to distinguish between models with typically sized samples when using only a single locus. Of course, much more power is achieved when loci are combined.

If y is random, then the death process for the coalescent is

$$n \to \begin{cases} n & \text{w.p. } 1 - \rho - \dfrac{n(n-1)}{4N} - \rho E\{(1-y)^n + ny(1-y)^{n-1}\} \\ n-1 & \text{w.p. } \dfrac{n(n-1)}{4N} + \rho\binom{n}{2}E\{y^2(1-y)^{n-2}\} \\ n-i & \text{w.p. } \rho\binom{n}{i+1}E\{y^{i+1}(1-y)^{n-i-1}\}, \quad i = 1 \ldots (n-1), \end{cases} \tag{19}$$

where all of the expectations are taken with respect to the distribution of y. If y is assumed to be β-distributed, then all of these expectations can be computed. The properties of this coalescent will be explored in a future article.

NEUTRAL EVOLUTION

As a stochastic force, pseudohitchhiking is very similar to genetic drift. Certain properties of population genetical models should be essentially independent of which of these two stochastic forces is present. The rate of neutral evolution, $k = u$, is one such property. Although the fact that the frequency of a neutral allele is a martingale under the pseudohitchhiking model makes this calculation entirely trivial, if we take a somewhat long-winded route through the derivation, it will give us some insights into the nature of neutral evolution in very large populations with hitchhiking.

Consider first the probability that a very rare neutral allele with frequency $x \approx 0$ never experiences a hitchhiking event. For simplicity, assume that y is constant. The fate of this rare allele over the first few generations is summarized in Table 1.

TABLE 1

The fate of a neutral allele

Generation	Frequency	Probability not chosen
0	x	$1 - x$
1	$x(1-y)$	$1 - x(1-y)$
2	$x(1-y)^2$	$1 - x(1-y)^2$
⋮	⋮	⋮
i	$x(1-y)^i$	$1 - x(1-y)^i$

The probability that the allele is never chosen is

$$\prod_{i=0}^{\infty}[1 - x(1-y)^i].$$

For very small x, this becomes

$$1 - x\sum_{i=0}^{\infty}(1-y)^i + \mathcal{O}(x^2) \sim 1 - x/y.$$

Thus, the probability that the allele does hitchhike is, asymptotically, x/y. It might seem surprising at first that an allele might not catch a ride at least once during the infinity of hitchhiking events in the population. The reason it does not is that its frequency is initially very low and then declines by a factor $(1 - y)$ with each subsequent event, making hitchhiking progressively less likely. If a very rare allele does catch a ride, its frequency will increase to

$$x^*(1-y) + y \approx y,$$

where x^* is its frequency in the generation when it first catches a ride.

The rate of fixation of neutral alleles can now be written in the suggestive form,

$$k = 2Nu\left(\frac{x}{y}\right)y = 2Nu\left(\frac{1}{2Ny}\right)y = u.$$

$2Nu$ is the mutational input each generation, the next term is the probability that a new allele gets a ride (x/y with $x = 1/2N$), and the final y is the probability of fixation of an allele whose frequency is y. When $N \to \infty$, we are left with a model where neutral alleles jump into the population at the rate u/y. When they do enter, their initial frequency is y.

It is instructive to see how the pseudohitchhiking model (in an infinite population) would have fared had it, rather than genetic drift, been the stochastic force used in Kimura and Ohta's (1971) classic article on the neutral theory. That article used two observations, $k \approx 10^{-7}$ and $F \approx 0.9$, to estimate two parameters, $u = 10^{-7}$ and $N_e = 2.5 \times 10^5$, for a species with one generation per year. Had the pseudohitchhiking model been used, the estimate of u would have remained the same. Using the obvious generalization of Equation 5,

$$F = \frac{1}{1 + 2u/\rho y^2},$$

we have $2u/\rho y^2 \approx 0.1$ or

$$\rho = \frac{2 \times 10^{-6}}{y^2}$$

for the rate of hitchhiking as a function of y. For example, $y = 0.2$, $\rho = 5 \times 10^{-5}$. Thus, the development of

the neutral theory would have worked just as well with pseudohitchhiking in an infinite population as with genetic drift in a finite population. Real populations may have both stochastic factors playing significant roles.

DISCUSSION

The possibility that stochastic effects from linked selection events are a more important stochastic force than genetic drift, *i.e.*, that

$$2\rho E\{y^2\} \gg \frac{1}{N},$$

has some very important implications:

Levels of polymorphism at neutral sites would be insensitive to population size. By contrast, when genetic drift is the main stochastic force, ssh $= 4Nu$ is linearly dependent on population size.

If, as seems plausible, $\rho E\{y^2\}$ is less variable between species than is N, then levels of variation should be relatively constant between species.

The frequency spectrum of alleles should be skewed from the neutral spectrum in a direction that leads to negative values of Tajima's D. The skew should be more extreme in regions of low recombination.

Assuming the correctness of the underlying model of selection, estimates of such quantities as Ns in large populations (*i.e.*, $N \gg 2\rho E\{y^2\}$) are actually estimates of $\sigma/(2\rho E\{y^2\})$ and, as such, should be much more similar across species than would be the case under genetic drift. If there were some correlation across species in the magnitude of positive and negative values of s, then this may even help explain why so many estimates of Ns are close to 1.

Genetic variation should be proportional to levels of recombination.

Ever since Lewontin raised the issue, population geneticists have wrestled with the apparent lack of sensitivity of levels of variation to the variation in population sizes between species and to the homogeneity of variation between species. Various solutions have been proposed, but few can readily account for the fact that the silent nucleotide site heterozygosities of most diploid species are within one order of magnitude of each other. We have before us a rather simple solution to the problem, and one that does not cause a radical change in our understanding of the stochastic dynamics of populations. Rather, it suggests a reinterpretation of the parameters of our stochastic models and a slight, though important, change in the nature of the coalescent.

Is linked selection a more important force than drift? In regions of low recombination, including mitochondria, the answer is quite possibly in the affirmative. What about regions of the genome with "normal" levels of recombination? In Drosophila, the site heterozygosity away from regions of low recombination is around $\pi = 0.006$. Using Equation 17, we have

$$\rho = \frac{2u}{\pi y^2} = 3.3 \times 10^{-7} y^{-2},$$

where we have used $u = 10^{-9}$ as a typical nucleotide substitution (and mutation) rate for a silent site. Such a rate of hitchhiking is not patently unreasonable even if the only source of hitchhiking events are amino acid substitutions within the same locus as the silent site. Within a locus, $r \approx 10^{-5}$ between distant sites. If the strength of selection acting on a typical substitution were around 10^{-3}, then $r/s \approx 0.01$, which would imply $y \approx 1$. Thus, $\rho \approx 10^{-7}$, which is typical for the rate of amino acid substitution in a coding region. However, if $y = 1$ and $N = \infty$, all of the mutations in a sample would be singletons, which is not observed, so some refinements of both the models and the parameters are needed before we accept the notion that linked selection may be a more important force than drift.

Of course, the real impact of hitchhiking involves events from many closely linked loci. The effects of hitchhiking events from more distant loci decrease with r. A full quantitative analysis of this combined effect will be discussed in a future publication as there are some complications stemming from the interactions of substitutions at closely linked loci on each other. Nonetheless, even this simple argument suggests that amino acid substitutions themselves could represent the hitchhiking agents required for our theory to be valid.

There is a much more intriguing, though largely unexplored, source of linked perturbations: meiotic drive. While there are some well-known and dramatic cases of male drive elements in natural populations such as segregation-distorter in Drosophila (Hiraizumi *et al.* 1960) and the *t*-allele in Mus (Lewontin and Dunn 1960), not much is known about segregation distortion in females, where, because only one of the four products of meiosis makes it into a gamete, it is much more likely to occur. The reason that so little is known about female meiotic drive is due to the technical problem of disassociating viability and drive effects of chromosomes. But, if particular chromosomes in nature were driven to higher frequency by a segregation advantage, then all of the alleles on those chromosomes would increase in frequency just as required in our model. Given the attraction of a drift-like stochastic force that is independent of population size, the possibility that chromosomes might experience transient drive should be seriously considered.

The stochastic effect of linked substitutions as captured in the pseudohitchhiking model is remarkably like genetic drift. The mean change in frequency of an allele is zero and the variance in the change is proportional to $x(1 - x)$. Should the domain of genetic drift be extended to include this new force or should it be

given another name entirely? When classifying the "factors of evolution," Wright (1955) used only the second-order moments. Thus, his definition of "random drift" does encompass pseudohitchhiking. Under Wright's classification, our title's phrase "genetic drift in an infinite population" makes perfect sense. If another name should prove useful, "genetic draft," as suggested to me by Bill Gilliland, is a good candidate as it is close to genetic drift and it continues the hitchhiking idiom by alluding to drafting to gain speed as practiced by bicyclists.

Other forms of linked selection will lead to different dynamics for neutral alleles. Some, like the TIM model (Takahata *et al.* 1975), will lower the heterozygosity and will skew the frequency spectrum to give $D < 0$ (Gillespie 1997). Thus, there is room for a great deal of additional work to describe the stochastic effects of linked selection in other contexts. Many of these effects may contribute even more to the divorce of genetic drift and population size.

I thank Dick Hudson, Chuck Langley, Ralph Haygood, Masaru Iizuka, and the Davis Evolution discussion Group for their many useful comments on this work. This article is dedicated to my friend and colleague Will Provine in recognition of his important contributions to the history of ideas in population genetics and for his tenacious campaign to consider a wider view of genetic drift. The research reported here was funded in part by National Science Foundation grant DEB-9527808.

LITERATURE CITED

Aguadé, M., and C. H. Langley, 1994 Polymorphism and divergence in regions of low recombination, pp. 67–76 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, London / New York.

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

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

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., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63:** 213–227.

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

Cox, D. R., 1962 *Renewal Theory*. Methuen & Co. Ltd., London.

Gillespie, J. H., 1993 Substitution processes in molecular evolution. I. Uniform and clustered substitutions in a haploid model. Genetics **134:** 971–981.

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

Gillespie, J. H., 1999 The role of population size in molecular evolution. Theor. Popul. Biol. **55:** 145–156.

Hiraizumi, Y., L. Sandler and J. F. Crow, 1960 Meiotic drive in natural populations of *Drosophila melanogaster*. III. Implications of the segregation-distorter locus. Evolution **14:** 433–444.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

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

Kimura, M., and T. Ohta, 1971 Protein polymorphism as a phase of molecular evolution. Nature **229:** 467–469.

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

Lewontin, R. C., and L. C. Dunn, 1960 The evolutionary dynamics of a polymorphism in the house mouse. Genetics **45:** 705–722.

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

Miyashita, N., 1990 Molecular and phenotypic variation of the *Zw* locus region in *Drosophila melanogaster*. Genetics **125:** 407–419.

Ohta, T., and H. Tachida, 1990 Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics **126:** 219–229.

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

Takahata, N., K. Ishii and H. Matsuda, 1975 Effect of temporal fluctuation of selection coefficient on gene frequency in a population. Proc. Natl. Acad. Sci. USA **72:** 4541–4545.

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

Watterson, G. A., 1982 Mutant substitutions at linked nucleotide sites. Adv. Appl. Prob. **14:** 206–224.

Watterson, G. A., 1984 Substitution times for mutant nucleotides. J. Appl. Prob. **19A:** 59–70.

Wright, S., 1955 Classification of the factors of evolution. Cold Spring Harbor Symp. Quant. Biol. **20:** 16–24D.

Communicating editor: R. R. Hudson