THEORETICAL POPULATION BIOLOGY 41, 237–254 (1992)

The Effect of Strongly Selected Substitutions on Neutral Polymorphism: Analytical Results Based on Diffusion Theory

WOLFGANG STEPHAN,* THOMAS H. E. WIEHE,*,† AND MARCUS W. LENZ*,†

* *Department of Zoology and* † *Program in Applied Mathematics, University of Maryland, College Park, Maryland 20742*

Received June 14, 1991

We derived analytical results for the reduction of the level of neutral polymorphism due to genetic hitchhiking using diffusion theory. In the case of a single strongly selected substitution, expected heterozygosity at a linked neutral locus is reduced by a factor $(2c/s)\,\alpha^{-2c/s}\Gamma(-2c/s, 1/\alpha)$, where s is the selective advantage of the favored allele, c is the recombination fraction between the neutral and selected locus, and $\alpha = 2Ns$, with N as the diploid population size. Γ denotes the incomplete gamma function. Using this result and assuming that at any one time at most one substitution is on its way to fixation, the effect of recurrent selected substitutions on expected heterozygosity can be approximated by $1/[1-\alpha(v/\rho)\,\lambda(1/\alpha, -2M/s)]$, where v and ρ are the expected numbers of selected substitutions and crossovers, respectively, per chromosome, per nucleotide, per generation, and M is the maximal recombination distance from the neutral locus that a selected mutation can be and still have a hitchhiking effect on the neutral locus. This expression involves a special function, $\lambda < 0$, which has a simple integral representation. The results are compared with those of deterministic population genetics and coalescent theory.

1. INTRODUCTION

When a selectively favored mutation occurs in a population and is subsequently fixed, the frequencies of polymorphisms at linked loci will be altered. This phenomenon is commonly referred to as the hitchhiking effect (Maynard Smith and Haigh, 1974). The hitchhiking effect has been analyzed by a number of authors (Kojima and Schaeffer, 1967; Maynard Smith and Haigh, 1974; Ohta and Kimura, 1975; Thomson, 1977; Kaplan, Hudson, and Langley, 1989). Two major scenarios have been modeled by these authors. Ohta and Kimura studied the effect on a neutral mutation which arises in the population while a selectively favored allele is on its way to fixation. In contrast, Maynard Smith and Haigh as well as Kaplan

0040-5809/92 $3.00

et al. analyzed the effect of a newly arising strongly selected allele on a (preexisting) polymorphic neutral locus. For strongly selected substitutions, the latter situation appears to be biologically more relevant, because strongly selected mutations go to fixation very rapidly. Maynard Smith and Haigh and Kaplan *et al.* came to the conclusion that the hitchhiking effect of strongly selected alleles leads to a reduction of the level of polymorphism at linked neutral loci. The main difference between their analyses is that Maynard Smith and Haigh present a deterministic model, whereas Kaplan *et al.* base their conclusions on coalescent theory. As a consequence, they find quantitative differences in the amount by which expected heterozygosity at the neutral locus is reduced due to hitchhiking. The deterministic approach led to some analytical results. The coalescent approach required extensive numerical calculations.

This paper considers also the consequences of a strongly selected substitution on preexisting linked neutral polymorphism and presents some analytical formulas. We study the following two situations: (1) the effect of a single substitution on expected heterozygosity, and (2) the case of recurrent substitutions where only one substitution event is allowed to happen at any one time. Our results are derived based on a modified version of Ohta and Kimura's (1975) moment analysis. Ohta and Kimura used expectations of functions of diffusion variables to model the hitchhiking dynamics. In their approach, the change in expected heterozygosity can be described by a set of ordinary differential equations for the first and second moments of the frequencies of the alleles at the neutral locus under consideration (instead of a partial differential equation). Because we chose to study a biological situation different than that of Ohta and Kimura, we had to modify their approach such that a strongly selected mutation arises in a population in which a neutral polymorphism preexists at a linked locus. We accounted for this situation by choosing different initial conditions for the set of ordinary differential equations for the moments.

With the advent of molecular population genetics and the continuing debate about the generality of the neutral theory of molecular evolution, the analysis of hitchhiking models found renewed interest. Recent data on intraspecific nucleotide polymorphism in *Drosophila* provide evidence for reduction of average nucleotide heterozygosity in regions of restricted crossing over per physical length, i.e., near centromeres and telomeres (Stephan and Langley, 1989; Aguadé, Miyashita, and Langley, 1989). The observed reduction appears to be in qualitative agreement with the hitchhiking effect. However, a quantitative analysis of such data requires more fully developed models of natural selection and genetic hitchhiking.

In contrast to Kaplan *et al.* (1989), we analyze the diffusion process directly, rather than the sample, because the diffusion approach can be generalized more readily than the coalescent approach to describe non-

neutral polymorphism. Since recent studies of molecular population genetics and evolution suggest that a considerable proportion of mutations at the molecular level is selected, such a generalization is desirable. In this (first) paper, however, we analyze the dynamics of hitchhiking alleles under the assumption of neutrality.

2. The Model

We study a two-locus model consisting of a selected and a linked neutral locus in order to investigate how much the level of heterozygosity at the neutral locus is changed by selected substitutions at the other locus. At the selected locus, the wildtype is denoted by b and the advantageous allele by B. Assuming no dominance, the fitnesses of the three genotypes bb, Bb, and BB can be assigned as $1, 1+s$, and $1+2s$, respectively. The alleles of the neutral locus are called A and a. We assume that a favorable mutation B arises in the population at time $t=0$ and is subsequently in the process of replacing allele b. This fixation process will alter the frequencies of alleles A and a at the linked neutral locus and may influence levels of heterozygosity in natural populations.

Our model assumes that two processes are acting in a random mating diploid population of (fixed) size N: random genetic drift and directional selection. As described above, selection operates only at the second locus. To analyze this two-locus, two-allele model, we follow Ohta and Kimura's (1975) treatment. Since a complete analysis of the underlying diffusion process is very difficult, we do not present here the full three-dimensional diffusion equation. Instead, we are separating the dynamics of the selected locus from that of the neutral locus. Furthermore, we consider only those sample paths in which the favored allele is going to fixation. This procedure is based on the assumption that selection is so strong that the dynamics of the selected locus is independent of the dynamics of the neutral locus and can be treated deterministically. The same approach was used by Kaplan *et al.* (1989). These authors point out that the above assumption is not stringent, because there is no hitchhiking effect when $\alpha = 2Ns$ is small. Our theoretical treatment is corroborated by Monte Carlo simulations.

If α is large, the stochastic frequency process $X(t)$ of the selected allele can be treated as deterministic as long as it stays away from the boundaries 0 and 1, because the trajectory of X is then sufficiently smooth. This approximation is based on mathematical arguments of Kurtz (1970). According to his theory, if $\varepsilon > 0$ and small ($\varepsilon \ll 1$), then with high probability

$$X(t) = x(t), \qquad t_\varepsilon \leqslant t \leqslant t_{1-\varepsilon}, \tag{1}$$

where

$$t_\varepsilon = \inf\{t : X(t) = \varepsilon\},$$
$$t_{1-\varepsilon} = x^{-1}(1-\varepsilon),$$

and $x(t)$ satisfies the differential equation

$$\frac{dx(t)}{dt} = sx(t)(1-x(t)), \qquad x(t_\varepsilon) = \varepsilon. \tag{2}$$

The solution of this differential equation is

$$x(t) = \frac{\varepsilon}{\varepsilon + (1-\varepsilon)\, e^{-s(t-t_\varepsilon)}}. \tag{3a}$$

It is convenient to introduce a new variable $\tau = t - t_\varepsilon$. The time it takes for the X process to go from ε to $1-\varepsilon$ is

$$\hat{\tau} = -2\ln(\varepsilon)/s. \tag{3b}$$

In order to describe the effect of the selected mutation on a linked neutral locus, Ohta and Kimura (1975) divide the population into two parts. One part consists of chromosomes carrying the advantageous mutation B, another one the disadvantageous allele b. Let p_1 be the frequency of allele A among chromosomes carrying the favorable mutation B, and p_2 the frequency of allele A among b-chromosomes. Note that these variables are different from the usual state space variables of two-locus, two-allele models. Furthermore, let $\phi(p_1, p_2, \tau)$ be the joint probability density function of p_1 and p_2 at time $\tau > 0$. Our goal is to compute the expectations of the frequencies p_1 and p_2 and their second-order moments p_1^2, $p_1 p_2$, and p_2^2 at time τ with respect to ϕ. For an arbitrary polynomial, f, of p_1 and p_2 we define these expectations as

$$E(f, \tau) = \int_0^1 \int_0^1 f(p_1, p_2)\, \phi(p_1, p_2, \tau)\, dp_1\, dp_2. \tag{4}$$

Because the frequency of allele A can be expressed by p_1 and p_2 as

$$p = p_1 x(\tau) + p_2(1 - x(\tau)), \tag{5}$$

this may allow us to calculate the expected heterozygosity at time τ. The general approach is to write down a set of ordinary differential equations

(ODEs) for the first- and higher-order moments of p_i, derived from the equation

$$\frac{d}{d\tau} E(f, \tau) = E(L(f), \tau), \tag{6}$$

where L is an appropriately defined differential operator (see below). In essence, this procedure has been used in the analysis of master and Fokker–Planck equations (reviewed by van Kampen, 1975). It was introduced into population genetics theory by Ohta and Kimura (1969). Because there is some confusion in the literature about the meaning of L, we have rederived (6) in Appendix 1. L is formally identical with the differential operator of the Kolmogorov backward equation. However, the variables are not those of the backward equation.

In the present case, Ohta and Kimura (1975) obtained the differential operator

$$L = \frac{p_1(1-p_1)}{4Nx(\tau)} \frac{\partial^2}{\partial p_1^2} + c(1-x(\tau))(p_2 - p_1) \frac{\partial}{\partial p_1} + \frac{p_2(1-p_2)}{4N(1-x(\tau))} \frac{\partial^2}{\partial p_2^2} + cx(\tau)(p_1 - p_2) \frac{\partial}{\partial p_2}, \tag{7}$$

where c is the recombination fraction between the two loci. They derived this operator noting that, in generation τ, the subpopulation carrying the advantageous mutation B consists of $2Nx(\tau)$ chromosomes, while the subpopulation carrying b has $2N(1-x(\tau))$ chromosomes. As mentioned above, the variables of the differential operator L are not those of the usual three-dimensional state space of two-locus, two-allele diffusion models. We note also that L contains only two variables, because the dynamics of the advantageous allele B is treated separately. Furthermore, we emphasize that the approach of Ohta and Kimura (1975) ignores most of the sample paths, i.e., the paths of the favored alleles which do not go to fixation.

To derive equations for the moments of the gene frequencies at the neutral locus, we choose f in (6) successively as p_1, p_2, p_1^2, $p_1 p_2$, and p_2^2. This leads to the following differential equations for the expectations of these quantities:

$$\frac{d}{d\tau} E(p_1) = c(1-x(\tau))\, E(p_2 - p_1), \tag{8a}$$

$$\frac{d}{d\tau} E(p_2) = cx(\tau)\, E(p_1 - p_2), \tag{8b}$$

$$\frac{d}{d\tau} E(p_1^2) = E\left(\frac{p_1(1-p_1)}{2Nx(\tau)} + 2c(1-x(\tau))\, p_1(p_2-p_1)\right), \tag{9a}$$

$$\frac{d}{d\tau} E(p_1 p_2) = E(c(1-x(\tau))\, p_2(p_2-p_1) + cx(\tau)\, p_1(p_1-p_2)), \tag{9b}$$

$$\frac{d}{d\tau} E(p_2^2) = E\left(\frac{p_2(1-p_2)}{2N(1-x(\tau))} + 2cx(\tau)\, p_2(p_1-p_2)\right). \tag{9c}$$

Two things are remarkable about these equations. First, the moment expansion breaks up. The equations for the first and second moments do not contain higher-order terms. Second, the first-order equations do not depend on second-order moments. The ODEs (8a)–(9c) are identical with those derived by Ohta and Kimura (1975). Ohta and Kimura consider the situation in which a neutral mutation appears in the population while the selected allele is on its way to fixation. Our goal is to study the effect of a newly arising favorable allele on polymorphism at the neutral locus. Therefore, we have to change the initial conditions for these equations. Furthermore, to obtain the expectation of any function f of p_1 and p_2, we distinguish the possibilities that the favorable mutation occurs on an A- or an a-carrying chromosome and calculate a weighted expectation as follows:

$$\bar{E}(f) = p_{20} E(f \mid p_{10} = 1) + (1 - p_{20})\, E(f \mid p_{10} = 0). \tag{10}$$

Here, p_{10} is the frequency of A-chromosomes (conditional on they also contain B) at time $t = 0$, when the favorable mutation arises, and p_{20} is the frequency of A-chromosomes (conditional on the presence of b) at time $t = 0$. The latter is equal to the frequency of allele A in the population at this time. Thus, the initial conditions for the above system of ODEs are determined by these two cases. Because these equations are formulated on the time scale of τ (instead of t), we have to make an approximation to get the corresponding initial conditions $p_{1\varepsilon}$ and $p_{2\varepsilon}$ on that scale. We simply assume that

$$p_{1\varepsilon} = p_{10} \qquad \text{and} \qquad p_{2\varepsilon} = p_{20}. \tag{11}$$

The second assumption is justified for large populations which we are considering here. In large populations, random genetic drift is unlikely to change the frequency of Ab-alleles in the short initial phase (Crow and Kimura, 1970, p. 383). The first assumption requires that no crossing involving B-chromosomes takes place between time $t = 0$ and t_ε. Unless recombination is infrequent, this assumption is harder to justify. We come back to this point at the end of the next section.

3. Effect of a Single Hitchhiking Event on Expected Heterozygosity

We derive analytical expressions for the expected heterozygosity $H_{1-\varepsilon}$ at the end of the selective phase, at time $t_{1-\varepsilon}$. Expected heterozygosity is defined by (10) with $f = 2p(1-p)$, where p is the frequency of allele A (see Eq. (5)). We begin by considering a simple case.

(i) *The Deterministic Limit.* The deterministic behavior is determined by the ODEs for the first-order moments, i.e., Eqs. (8a) and (8b). These equations are identical with Maynard Smith and Haigh's Eqs. (19) and (20), if in the latter ones $s \ll 1$ is assumed. The solutions of these ODEs can readily be found as follows:

$$E(p_1) = p_{1\varepsilon} - c(p_{1\varepsilon} - p_{2\varepsilon}) \int_0^\tau \frac{(1-\varepsilon)\, e^{-(s+c)\tau'}}{\varepsilon + (1-\varepsilon)\, e^{-s\tau'}}\, d\tau', \tag{12a}$$

$$E(p_2) = p_{2\varepsilon} + c(p_{1\varepsilon} - p_{2\varepsilon}) \int_0^\tau \frac{\varepsilon e^{-c\tau'}}{\varepsilon + (1-\varepsilon)\, e^{-s\tau'}}\, d\tau'. \tag{12b}$$

At the end of the selective phase, at time $t_{1-\varepsilon}$ (or, equivalently, at $\hat{\tau} = -2\ln(\varepsilon)/s$), the frequency of allele A is

$$p = p_1 + \varepsilon(p_2 - p_1). \tag{13}$$

Therefore, at time $\hat{\tau}$, $E(p) = E(p_1) + (p_{2\varepsilon} - p_{1\varepsilon})\, \varepsilon^{1+2c/s}$. Because in the deterministic limit, $E(p) = p$, we find for average heterozygosity at the end of the selective phase

$$\frac{H_{1-\varepsilon}}{2p_{2\varepsilon}(1-p_{2\varepsilon})} = cI(\hat{\tau})(2 - cI(\hat{\tau})) + O(\varepsilon^{1+3c/s}), \tag{14a}$$

where

$$I(\tau) = \int_0^\tau \frac{e^{-(c+s)\tau'}}{\varepsilon + e^{-s\tau'}}\, d\tau'. \tag{14b}$$

Under the assumptions (11), we have at the beginning of the selective phase at time t_ε, $\bar{E}(p) = p_{2\varepsilon}$, and hence $H_\varepsilon = 2p_{2\varepsilon}(1-p_{2\varepsilon})$. Therefore, formulas (14a) and (14b) express the reduction of average heterozygosity due to a single hitchhiking event. We were able to derive a simpler expression by approximating the integral $I(\tau)$. $I(\tau)$ describes the time-dependent behavior of $E(p_1)$ (see (12a)). Equation (8a) suggests that $E(p_1)$ changes only for $\tau < \frac{1}{2}\hat{\tau}$, so long as the frequency of the selected allele B is low.

Hence, $I(\hat{\tau}) \approx I(\frac{1}{2}\hat{\tau})$. Because the frequency of allele B is low for $\tau < \frac{1}{2}\hat{\tau}$, we may replace the denominator in (14b) by $e^{-s\tau'}$. This leads to

$$cI(\hat{\tau}) \approx 1 - \varepsilon^{c/s} \tag{14c}$$

and

$$\frac{H_{1-\varepsilon}}{2p_{2\varepsilon}(1-p_{2\varepsilon})} \approx 1 - \varepsilon^{2c/s}. \tag{14d}$$

For $\varepsilon = 1/2N$, (14c) corresponds to Maynard Smith and Haigh's (1974) deterministic value (see their formula (23), with s replaced by $2s$; in their model, the selective advantage of allele B is $\frac{1}{2}s$).

(ii) *The More General Case.* Equation (13) says that, at that time $t_{1-\varepsilon}$, $p \approx p_1$. Therefore, we attempt to derive a differential equation for $E(p_1(1-p_1))$. Subtracting Eq. (9a) from (8a) and substituting $\Delta(\tau) = (p_{1\varepsilon} - p_{2\varepsilon})\, e^{-c\tau}$ for $E(p_1 - p_2)$ leads to the following equation:

$$\frac{d}{d\tau} E(p_1(1-p_1)) + E(p_1(1-p_1))\left[\frac{1}{2Nx(\tau)} + 2c(1-x(\tau))\right]$$
$$= c(1-x(\tau))[2E(p_1(1-p_2)) - \Delta(\tau)]. \tag{15}$$

The right-hand side of this equation needs further attention, because it contains the expectation of $p_1 p_2$, which couples Eq. (15) to the other second-moment equations. Equations (8a) and (9a) suggest that expected heterozygosity changes markedly only for $\tau < \frac{1}{2}\hat{\tau}$, so long as the frequency of the selected allele B is low. The frequency of the Ab allele stays almost constant during this time period at $p_{2\varepsilon}$, its initial value. This is indicated by Eq. (8b). Therefore, we used the following approximation for $\tau < \frac{1}{2}\hat{\tau}$: $E(p_1 p_2) \approx p_{2\varepsilon} E(p_1)$. We have tested this approximation by integrating Eqs. (8a)–(9c) numerically. Using this approximation and taking the weighted expectation according to definition (10), the right-hand side of (15) can be simplified to $2c(1-x(\tau))\, p_{2\varepsilon}(1-p_{2\varepsilon})$, so long as $\tau < \frac{1}{2}\hat{\tau}$. Hence, in general we find

$$\frac{d}{d\tau} \bar{E}(p_1(1-p_1)) + \bar{E}(p_1(1-p_1))\left[\frac{1}{2Nx(\tau)} + 2c(1-x(\tau))\right]$$
$$= 2c(1-x(\tau))\, \bar{E}(p_1(1-p_2)); \tag{16a}$$

for $\tau < \frac{1}{2}\hat{\tau}$, the right-hand side of (16a) can be approximated by

$$2c(1-x(\tau))\, p_{2\varepsilon}(1-p_{2\varepsilon}). \tag{16b}$$

This differential equation can be solved approximately (see Appendix 2).

The reduction in expected heterozygosity at the end of the selected phase is then

$$\frac{H_{1-\varepsilon}}{2p_{2\varepsilon}(1-p_{2\varepsilon})}=\frac{2c}{s}\alpha^{-2c/s}\Gamma\left(-\frac{2c}{s},\frac{1}{\alpha},\frac{1}{\alpha\varepsilon}\right), \tag{17}$$

where $\Gamma(\gamma, y, z)$ is a generalized incomplete gamma function defined by $\Gamma(\gamma, y)-\Gamma(\gamma, z)$ (Gradshteyn and Ryzhik, 1980, formula (8.350.2)). In Table I, we compare this result to the numerical solution for $2N=10^8$, $\alpha=10^5$, $\varepsilon=100/2N$, and different values of c. Also shown in Table I are the numerical results obtained from the coalescent approach (Kaplan *et al.*, 1989). Furthermore, formula (17) is compared with the deterministic result (Eqs. (14a) and (14d)). As expected, the deterministic approach overestimates the reduction in heterozygosity for small ε, but agrees well with the more general result for $\varepsilon \geqslant 5/\alpha$ (results not shown).

If $\varepsilon \leqslant 1/\alpha$, then

$$\Gamma\left(-\frac{2c}{s},\frac{1}{\alpha},\frac{1}{\alpha\varepsilon}\right)\approx\Gamma\left(-\frac{2c}{s},\frac{1}{\alpha}\right), \tag{18}$$

TABLE I

Reduction of Heterozygosity Due to a Single Hitchhiking Event

$-\log_{10}(c/s)$	Eq. (14a)	Eq. (14d)	Eq. (17)	KHL	Runge–Kutta
3.0	0.027250	0.027253	0.021631	0.021637	0.021626
2.8	0.042839	0.042847	0.034062	0.034079	0.034050
2.6	0.067033	0.067052	0.053437	0.053477	0.053409
2.4	0.104120	0.104167	0.083341	0.083376	0.083279
2.2	0.159879	0.159989	0.128786	0.128580	0.128650
2.0	0.241173	0.241422	0.196182	0.195660	0.195883
1.8	0.354090	0.354624	0.292346	0.291868	0.291707
1.6	0.499418	0.500456	0.421466	0.419731	0.420186
1.4	0.665391	0.667132	0.579071	0.576768	0.576761
1.2	0.822766	0.825075	0.744880	0.740796	0.741370
1.0	0.934787	0.936904	0.883579	0.878784	0.879500
0.8	0.986376	0.987465	0.965524	0.961843	0.962341
0.6	0.998804	0.999032	0.994531	0.992985	0.993085

Note. Populations $2N=10^8$; $\alpha=10^5$; $\varepsilon=10^{-6}$. For explanation of Eq. (14a), (14d), and (17) see the text. The data in the last column have been calculated by solving the system of five ordinary differential equations (Eqs. (8a)–(9c)) by means of the Runge–Kutta method. The data in column KHL show linearly interpolated P_{22}-data from Kaplan *et al.* (1989). The numerical values have been kindly provided by C. H. Langley.

so that the reduction in expected heterozygosity is only weakly dependent on ε, i.e.,

$$\frac{H_{1-\varepsilon}}{2p_{2\varepsilon}(1-p_{2\varepsilon})} \approx \frac{2c}{s}\alpha^{-2/s}\Gamma\left(-\frac{2c}{s}, \frac{1}{\alpha}\right). \tag{19}$$

Dr. J. Gillespie has examined this formula by computer simulations using a Wright–Fisher model. In these simulations, the initial frequency of the selected allele is $1/2N$, and that of the neutral allele A is p_0. Hence, the initial phase ($t < t_\varepsilon$) as well as the phase prior to fixation ($t > t_{1-\varepsilon}$), which we do not treat here, are included in the simulations. Those selected mutations that go to fixation reduce heterozygosity at the neutral locus by the amounts shown in Table II. The simulation results are compared with (19). The simulations agree well with the expected results, if α is sufficiently large, as we have assumed throughout this derivation. However, discrepancies become apparent for smaller α, suggesting that the present analysis holds only asymptotically. The simulations also indicate that the reduction in expected heterozygosity is independent of the initial frequency of allele A. This is consistent with our approximation result which shows that the right-hand side of Eq. (17) is independent of $p_{2\varepsilon}$.

TABLE II

Reduction of Heterozygosity Due to a Single Hitchhiking Event: Comparison of Theoretical and Simulations Results

	Simulation				
$-\log_{10}(c/s)$	$p_0 = 0.5$		$p_0 = 0.1$		Expected
			$\alpha = 2 \times 10^3$		
3.0	0.0124	(0.0034)	0.0128	(0.0016)	0.0139
2.0	0.1208	(0.0123)	0.1302	(0.0075)	0.1308
1.0	0.7315	(0.0363)	0.7378	(0.0078)	0.7456
0.0	0.9919	(0.0130)	0.9910	(0.0006)	0.9990
			$\alpha = 2 \times 10^2$		
3.0	0.0051	(0.0021)	0.0050	(0.0025)	0.0094
2.0	0.0928	(0.0123)	0.0944	(0.0192)	0.0899
1.0	0.5342	(0.0139)	0.5240	(0.0367)	0.5977
0.0	0.9320	(0.0440)	0.9115	(0.0306)	0.9902

Note. Population size $N = 10^4$. Column "expected" shows values calculated from Eq. (19). The values in parentheses represent the standard error for the different initial frequencies $p_0 = 0.5$, $p_0 = 0.1$, respectively. For each parameter set, mean and standard error have been calculated based on 400 substitution events.

4. Effect of Recurrent Selected Substitutions on Heterozygosity

We assume that substitutions occur randomly along the chromosome according to a time-homogeneous Poisson process, and that at any one time at most only one substitution (linked to the neutral locus under consideration) is on its way to fixation. The reduction of average heterozygosity at the neutral locus is then given by the probability that the neutral locus escapes hitchhiking under recurring substitutions. Let v be the expected number of selected substitutions per nucleotide site, per chromosome, per generation, and let ρ be the expected number of cross-overs. Furthermore, let k_h be the rate at which selected substitutions occur that drag the neutral locus to fixation. To calculate k_h we assume that the chromosome is continuous. The rate at which substitutions occur between m and $m+dm$ nucleotide sites away from the neutral locus, and take the neutral locus along, is $2Nv(1-h(\rho m))\,dm$; $h(\rho m)$ denotes the right-hand side of (14a), (14d), (17), or (19) with ρm for c. Averaging this expression over m and scaling back to the original variable c leads to

$$k_h = 4N\frac{v}{\rho}\int_0^M (1-h(c))\,dc. \tag{20}$$

Here, M is the maximum of the recombinational distance from the neutral locus over which selected substitutions are assumed to occur. (This corresponds to a maximum physical distance of M/ρ.) Since we allow in this model only a single substitution event at any one time, maximum values for M and v/ρ need to be specified. This can be done according to Kaplan *et al.* (1989).

We now consider a sequence of substitution events occurring in $[0, M/\rho]$. The probability that the neutral locus is dragged along by one of these substitutions is $k_h/(1+k_h)$, and, conversely, that it escapes is $1/(1+k_h)$. Therefore, the effect of recurrent substitutions on expected heterozygosity is

$$H = \frac{H_{\text{neu}}}{1+k_h}, \tag{21}$$

where H_{neu} denotes the neutral value. This corresponds to formula (18) of Kaplan *et al.* (1989).

The integral in (20) can be readily evaluated in the following case. Using the series representation of the incomplete gamma function (Gradshteyn and Ryszik, 1980, formula (8.354.2)) and keeping only the lowest-order

terms (because α is assumed to be very large), we find for the right-hand side of (19)

$$\frac{H_{1-\varepsilon}}{2p_{2\varepsilon}(1-p_{2\varepsilon})} \approx 1-\alpha^{-2c/s}\Gamma\left(1-\frac{2c}{s}\right). \tag{22}$$

For the parameter values used in Table I, this approximation deviates from (17) by less than 2 %, even when $\varepsilon = 1/\alpha$. The integral over $\alpha^{-2c/s}\Gamma(1-2c/s)$ can be expressed by a special function $\lambda(\cdot,\cdot)$ (Gradshteyn and Ryszik, 1980, formula (9.640.5)) such that

$$k_h \approx -\alpha\frac{v}{\rho}\lambda\left(\frac{1}{\alpha}, -\frac{2M}{s}\right). \tag{23a}$$

For numerical calculation the integral representation (Gradshteyn and Ryszik, 1980, formula (4.359.1))

$$-\lambda(\beta,-\gamma) = \beta\int_0^\infty e^{-\beta x}\frac{1-x^{-\gamma}}{\ln x}\,dx, \qquad 0<\beta, 0<\gamma<1, \tag{23b}$$

can be used since M is chosen such that $2M/s < 1$ (see Table 2 of Kaplan *et al.*, 1989; in our notation, their M value has to be divided by $2N$). Using Eq. (23), the reduction of heterozygosity below the neutral level can be calculated. The results agree well with those summarized in Fig. 4 of Kaplan *et al.*

5. Discussion

In this study we have reformulated the approach of Ohta and Kimura (1975) in order to describe the effect of strongly selected substitutions on a neutral polymorphism. We obtained analytical results on the effect of a single substitution on expected heterozygosity and also for the recurrent case in which at most only one substitution event occurs at any one time. The reduction of heterozygosity due to hitchhiking had previously been studied by Maynard Smith and Haigh (1974) using a deterministic model and by Kaplan *et al.* (1989) based on coalescent theory. Our results agree well with those of the coalescent approach. However, as shown in Table I, the deterministic approximation may lead to a considerable overestimation of the reduction in expected heterozygosity. Our analytical results are also in agreement with the simulations carried out by Dr. J. Gillespie, if $\alpha = 2Ns$ is sufficiently large. Comparison with the simulations suggests that the present approach holds only asymptotically, for large α.

Reduction of average nucleotide heterozygosity by approximately one order of magnitude has been observed in natural populations of *Drosophila*. For instance, Aguadé *et al.* (1989) surveyed a 100-kb region located near the telomere of the X chromosome in *D. melanogaster* encompassing the *yellow-achaete-scute* complex and found an approximately 10-fold reduction in average nucleotide heterozygosity compared with other euchromatic regions. The rate of crossing over per physical length in this region is at least 20-fold lower than in euchromatin. Reduced variation is also found at the tip of the X chromosome in *D. simulans* (Begun and Aquadro, 1991). Similar observations have been made at the base of the X chromosome in *D. ananassae* (Stephan and Langley, 1989), where recombination is also heavily restricted. Because recombination near centromeres and telomeres is suppressed over large distances, low levels of variation have been found over relatively large segments of DNA. In regions of intermediate and high recombination rates, the hitchhiking effect should be less apparent. In those parts of the chromosome, regions of reduced variation can be expected to be short. Kreitman and Hudson (1991) have suggested that the deficiency of variation at the *Adh-dup* locus may be explained by hitchhiking. The low level of polymorphism observed by Lange, Langley, and Stephan (1990) in the *Mtn* region in *D. melanogaster* may also be attributable to a local hitchhiking effect. However, in the latter two cases the reduction in nucleotide heterozygosity is not statistically significant.

The present approach is only part of a theory of directional selection and genetic hitchhiking, because only the effect on neutral polymorphism is considered. Given that there is likely to be a spectrum of selected mutations, a more general theory of the evolutionary dynamics of hitchhiking alleles is desirable. Various questions need to be addressed by such a theory. It is conceivable that weakly selected alleles interfere with the strongly selected ones and thus reduce the efficiency of the hitchhiking effect. A weakly selected allele on its way to fixation stays in low frequency for a long time. Therefore, strongly selected mutations are likely to occur in repulsion. However, in the time intervals between successive substitution events of strongly favored alleles, the weakly selected mutation may increase in frequency (without making it to fixation) and by this process drag along neutral (or nearly neutral) polymorphisms. This would lead to an increase of heterozygosity which could transiently stay at a relatively high level and obscure the hitchhiking effect. Ohta and Kimura's (1975) results could be interpreted in that way, although, given the approach they chose, they certainly did not model the dynamics of a weakly selected substitution. Similarly, a theory which may ultimately lead to a correct understanding of levels of molecular variation in natural populations needs to include the effects of strongly selected mutations on (slightly) deleterious alleles (e.g., alleles carrying transposable elements). Progress in this direction

can be made based on the diffusion approach. The advantage of the diffusion approach is that it will generalize if the assumption of neutrality for the hitchhiking alleles is dropped.

APPENDIX 1: DERIVATION OF EQUATION (6)

In the case of a one-dimensional diffusion process, Eq. (6) takes the form

$$\frac{d}{dt} E_t(f) = E_t\left(\mu(p,t)\frac{d}{dp} f + \frac{1}{2}\sigma^2(p,t)\frac{d^2}{dp^2} f\right), \tag{A.1}$$

where $\mu(p, t)$ and $\sigma^2(p, t)$ denote the infinitesimal drift and diffusion parameters, respectively, defined in Karlin and Taylor (1981, p. 159). To derive (A.1), denote the density function of the gene frequency distribution by $\phi(p, t)$ and define a functional acting on a twice continuously differentiable function f by

$$E_t(f) := \int_0^1 f(p)\,\phi(p, t)\,dp \equiv E(f, t). \tag{A.2}$$

In the following we omit the integration boundaries. All integrals are to be taken between 0 and 1. $\phi(p, t)$ obeys the equation

$$\phi(p, t+s) = \int P(p, t+s \mid \xi, t)\,\phi(\xi, t)\,d\xi, \tag{A.3}$$

where $P(p, t+s \mid \xi, t)$ denotes the conditional probability that the process is in state p at time $t+s$, given that it was in state ξ at time t ($t>0, s\geqslant 0$). Using (A.3) and expanding f in a Taylor series around ξ, we find (neglecting terms of order $\geqslant 3$)

$$\begin{aligned} E_{t+h}(f) &= \int f(p)\,\phi(p, t+h)\,dp = \int f(p) \int P(p, t+h \mid \xi, t)\,\phi(\xi, t)\,d\xi\,dp \\ &= \int \left\{ \int f(\xi)\,P(p, t+h \mid \xi, t)\,dp + \int f'(\xi)(p-\xi)\,P(p, t+h \mid \xi, t)\,dp \right. \\ &\quad \left. + \int \frac{1}{2} f''(\xi)(p-\xi)^2\,P(p, t+h \mid \xi, t)\,dp \right\} \phi(\xi, t)\,d\xi. \end{aligned} \tag{A.4}$$

Equation (A.4) can be rewritten using

$$\int \left\{ \int f(\xi)\,P(p, t+h \mid \xi, t)\,dp \right\} \phi(\xi, t)\,d\xi = E_t(f). \tag{A.5}$$

Subtracting (A.5) from (A.4) and dividing both sides of the resulting equation by h, we obtain

$$\frac{1}{h}(E_{t+h}(f)-E_t(f))=E_t\left(f'\int\frac{1}{h}(p-\xi)\,P(p,\,t+h\,|\,\xi,\,t)\,dp\right)$$

$$+E_t\left(\frac{1}{2}f''\int\frac{1}{h}(p-\xi)^2\,P(p,\,t+h\,|\,\xi,\,t)\,dp\right). \qquad \text{(A.6)}$$

Assuming that

$$\int\frac{1}{h}(p-\xi)\,P(p,\,t+h\,|\,\xi,\,t)\,dp\to\mu(\xi,\,t), \qquad \text{as} \quad h\to 0$$

and

$$\int\frac{1}{h}(p-\xi)^2\,P(p,\,t+h\,|\,\xi,\,t)\,dp\to\sigma^2(\xi,\,t), \qquad \text{as} \quad h\to 0,$$

and assuming that it is justified to pass to the limit under the integral, (A.6) leads to (A.1),

$$\frac{d}{dt}E_t(f)=E_t\left(\mu(p,\,t)\frac{d}{dp}f+\frac{1}{2}\sigma^2(p,\,t)\frac{d^2}{dp^2}f\right),$$

where variables are switched back from ξ to p. In the case of a n-dimensional diffusion process, an analogous equation can be obtained using the same methods. In this case, the infinitesimal parameters of the diffusion are given by

$$\lim_{h\to 0}\int\frac{1}{h}(p_i-\xi_i)\,P(p,\,t+h\,|\,\xi,\,t)\,dp=\mu_i(\xi,\,t), \qquad \text{(A.7)}$$

$$\lim_{h\to 0}\int\frac{1}{h}(p_i-\xi_i)(p_j-\xi_j)\,P(p,\,t+h\,|\,\xi,\,t)\,dp=\sigma_{ij}(\xi,\,t), \qquad i\neq j, \qquad \text{(A.8)}$$

and

$$\lim_{h\to 0}\int\frac{1}{h}(p_i-\xi_i)^2\,P(p,\,t+h\,|\,\xi,\,t)\,dp=\sigma_i^2(\xi,\,t), \qquad i,\,j=1,\,...,\,n. \qquad \text{(A.9)}$$

Here, we also assume that these limits exist in a "sufficiently regular" sense. The general form of (A.1) then becomes

$$\frac{d}{dt}E_t(f) = E_t\left(\sum_{i=1}^{n}\mu_i(p,t)\frac{\partial}{\partial p_i}f + \frac{1}{2}\sum_{i=1}^{n}\sigma_i^2(p,t)\frac{\partial^2}{\partial p_i^2} + \sum_{i,j=1}^{n}\sigma_{ij}(p,t)\frac{\partial^2}{\partial p_i\,\partial p_j}f\right). \quad \text{(A.10)}$$

APPENDIX 2: DERIVATION OF EQUATION (17)

The ODE (16) has the general form

$$\frac{d}{d\tau}h(\tau) + h(\tau)\,P(\tau) = Q(\tau) \quad \text{(A.11)}$$

with

$$h(\tau) = \bar{E}(p_1(1-p_1))/[p_{2\varepsilon}(1-p_{2\varepsilon})],$$

$$P(\tau) = \frac{1}{2Nx(\tau)} + 2c(1-x(\tau)),$$

$$Q(\tau) = 2c(1-x(\tau))\,q(\tau),$$

and

$$q(\tau) = \bar{E}(p_1(1-p_2))/[p_{2\varepsilon}(1-p_{2\varepsilon})].$$

The general solution of the homogeneous ODE is

$$h_{\text{hom}}(\tau) = C\exp\left(-\int_0^\tau P(\tau')\,d\tau'\right). \quad \text{(A.12)}$$

The particular solution which satisfies the inhomogeneous ODE (A.11) and the initial condition $h(0)=0$ (see the definition of the weighted expectation) can be found by variation of parameters as

$$h(\tau) = C(\tau)\,h_{\text{hom}}(\tau), \quad \text{(A.13)}$$

where

$$C(\tau) = \int_0^\tau Q(\tau')\exp\left(\int_0^{\tau'} P(\tau'')\,d\tau''\right)d\tau'. \quad \text{(A.14)}$$

The integral in (A.12) can be immediately evaluated. Because of $\varepsilon \ll 1$ and large population size, we obtain

$$\int_0^\tau P(\tau')\,d\tau' \approx \frac{1}{\alpha\varepsilon}(1-e^{-s\tau}) - \frac{2c}{s}\ln(\varepsilon + e^{-s\tau}), \tag{A.15}$$

and, at $\tau = \hat{\tau}$,

$$\exp\left(-\int_0^\tau P(\tau')\,d\tau'\right) \approx \varepsilon^{2c/s} e^{-1/\alpha\varepsilon}. \tag{A.16}$$

Using (A.15), $C(\hat{\tau})$ can be written as

$$C(\hat{\tau}) = 2c\exp\left(\frac{1}{\alpha\varepsilon}\right)\int_0^{\hat{\tau}} \frac{e^{-s\tau}}{\varepsilon + e^{-s\tau}} \exp\left(-\frac{1}{\alpha\varepsilon}e^{-s\tau} - \frac{2c}{s}\ln(\varepsilon + e^{-s\tau})\right) q(\tau)\,d\tau. \tag{A.17}$$

$C(\hat{\tau})$ has a form similar to that of $I(\hat{\tau})$ (see (14b)). Therefore, we may apply the same approximation technique used for calculating $I(\hat{\tau})$ and integrate only from 0 to $\frac{1}{2}\hat{\tau}$. For $\tau < \frac{1}{2}\hat{\tau}$, $\varepsilon + e^{-s\tau}$ can be approximated by $e^{-s\tau}$ and q by 1 (see text). This leads to

$$C(\hat{\tau}) \approx 2c\exp\left(\frac{1}{\alpha\varepsilon}\right)\int_0^{\hat{\tau}/2} \exp\left(-\frac{1}{\alpha\varepsilon}e^{-s\tau} + 2c\tau\right) d\tau. \tag{A.18}$$

Introducing a new variable $z = (1/\alpha\varepsilon)\,e^{-s\tau}$, the integral in (A.18) can be expressed by means of the incomplete gamma function. By combining this result with (A.16), we obtain

$$h(\hat{\tau}) \approx \frac{2c}{s}\alpha^{-2c/s}\left[\Gamma\left(-\frac{2c}{s}, \frac{1}{\alpha}\right) - \Gamma\left(-\frac{2c}{s}, \frac{1}{\alpha\varepsilon}\right)\right]. \tag{A.19}$$

ACKNOWLEDGMENTS

We are very grateful to John Gillespie for his computer simulations, which improved the quality of the present analysis considerably. We also thank Chuck Langley for providing us with the P_{22} data, which allowed us to compare our results with those of coalescent theory. Sam Mitchell helped with the numerical calculations. John Gillespie made helpful suggestions for improving the presentation of this paper. W. S. acknowledges the receipt of a 1990 Summer Research Award from the General Research Board of the University of Maryland.

REFERENCES

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

BEGUN, D. J., AND AQUADRO, C. F. 1991. Natural selection and recombination in *Drosophila*: Genetic hitchhiking in the distal portion of the *X* chromosome, *Genetics* **129**, 1147–1158.

CROW, J. F., AND KIMURA, M. 1970. "An Introduction to Population Genetics Theory," Harper & Row, New York.

GRADSHTEYN, I. S., AND RYZHIK, I. M. 1980. "Table of Integrals, Series, and Products," Academic Press, New York.

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

KARLIN, S., AND TAYLOR, H. M. 1981. "A Second Course in Stochastic Processes," Academic Press, New York.

KOJIMA, K. I., AND SCHAEFFER, H. E. 1967. Survival process of linked genes, *Evolution* **21**, 518–531.

KREITMAN, M., AND HUDSON, R. R. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence, *Genetics* **127**, 565–582.

KURTZ, T. G. 1970. Solutions of ordinary differential equations as limits of pure jump Markov processes, *J. Appl. Prob.* **7**, 49–58.

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

MAYNARD SMITH, J., AND HAIGH, J. 1974. The hitchhiking effect of a favorable gene, *Genet. Res.* **23**, 23–35.

OHTA, T., AND KIMURA, M. 1969. Linkage disequilibrium due to random genetic drift, *Genet. Res.* **13**, 47–55.

OHTA, T., AND KIMURA, M. 1975. The effect of a selected linked locus on heterozygosity of neutral alleles (the hitchhiking effect), *Genet. Res.* **25**, 313–326.

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

THOMSON, G. 1977. The effect of a selected locus on linked neutral loci, *Genetics* **85**, 753–788.

VAN KAMPEN, N. G. 1975. The expansion of the master equation, *in* "Advances in Chemical Physics" (I. Prigogine and S. A. Rice, Eds.), Wiley, New York.

Printed in Belgium