Analysis of a Genetic Hitchhiking Model, and Its Application to DNA Polymorphism Data from *Drosophila melanogaster*[1]

*Thomas H. E. Wiehe**,† *and Wolfgang Stephan**

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

Begun and Aquadro have demonstrated that levels of nucleotide variation correlate with recombination rate among 20 gene regions from across the genome of *Drosophila melanogaster*. It has been suggested that this correlation results from genetic hitchhiking associated with the fixation of strongly selected mutants. The hitchhiking process can be described as a series of two-step events. The first step consists of a strongly selected substitution wiping out linked variation in a population; this is followed by a recovery period in which polymorphism can build up via neutral mutations and random genetic drift. Genetic hitchhiking has previously been modeled as a steady-state process driven by recurring selected substitutions. We show here that the characteristic parameter of this steady-state model is $\alpha\nu$, the product of selection intensity ($\alpha = 2Ns$) and the frequency of beneficial mutations ν (where N is population size and s is the selective advantage of the favored allele). We also demonstrate that the steady-state model describes the hitchhiking process adequately, unless the recombination rate is very low. To estimate $\alpha\nu$, we use the data of DNA sequence variation from 17 *D. melanogaster* loci from regions of intermediate to high recombination rates. We find that $\alpha\nu$ is likely to be $>1.3 \times 10^{-8}$. Additional data are needed to estimate this parameter more precisely. The estimation of $\alpha\nu$ is important, as this parameter determines the shape of the frequency distribution of strongly selected substitutions.

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, which is commonly referred to as the "hitchhiking effect," was first analyzed by Maynard Smith and Haigh (1974). Aguadé et al. (1989) and Stephan and Langley (1989) used this effect as an indirect means of detecting the action of directional selection at the DNA level. They observed reduced levels of molecular variation in chromosomal regions in which the rate of meiotic recombination is severely restricted, i.e., near centromeres and telomeres. Their results are in qualitative agreement with hitchhiking models (Kaplan et al. 1989; Stephan et al. 1992) which have been developed on the basis of Maynard Smith and Haigh's (1974) suggestion.

In the past 5 years, many studies of DNA sequence variation in *Drosophila* have focused on the detection of natural selection at the DNA level by using the hitchhiking effect. The basic observation of reduced polymorphism in regions of restricted recom-

1. Key words: DNA sequence variation, genetic hitchhiking, advantageous mutations, molecular population genetics, *Drosophila*.

Address for correspondence and reprints: Wolfgang Stephan, Department of Zoology, University of Maryland, College Park, Maryland 20742.

Mol. Biol. Evol. 10(4):842–854. 1993.

0737-4038/93/1004-0008$02.00

bination, such as *y-ac-sc,* has been contested by several authors (Beech and Leigh Brown 1989; Eanes et al. 1989; Macpherson et al. 1990), mainly because the estimates of average nucleotide heterozygosity vary greatly between population samples. However, using newly developed statistical techniques and larger data sets, recent population surveys have confirmed the original observation of reduced variation in regions of restricted recombination in three *Drosophila* species—*D. melanogaster* (Berry et al. 1991; Martín-Campos et al. 1992; Langley et al. 1993), *D. simulans* (Begun and Aquadro 1991; Langley et al. 1993) and *D. ananassae* (Stephan and Mitchell 1992). Furthermore, Begun and Aquadro (1992), using 20 gene regions from across the genome of *D. melanogaster,* have demonstrated that levels of variation correlate with recombination rate, suggesting that hitchhiking occurs over a large fraction of the *Drosophila* genome.

Crucial questions concern the strength and frequency with which selective substitutions have to be postulated in order to explain the observations. In this study, we examine whether these questions can be addressed on the basis of our previous hitchhiking model.

Modeling the Hitchhiking Process

The hitchhiking process can be described as a series of two-step events, where each two-step event consists of (1) a strongly selected substitution wiping out standing linked variation in a population and (2) a neutral recovery period in which polymorphism can build up via drift and mutation. Both Kaplan et al. (1989) and Stephan et al. (1992) modeled the hitchhiking process as a steady-state process with recurring selected substitutions, neutral mutation, random genetic drift, and recombination. In this section, we study some properties of this model. Furthermore, by introducing a very simple model of the transient recovery period, we demonstrate the limitations of the steady-state approach for regions of very low recombination rates.

Effect of a Single Hitchhiking Event

We begin by describing the effect of a single hitchhiking event on neutral polymorphism. We consider a two-locus model consisting of a selected locus and a linked neutral locus. Suppose, in a randomly mating diploid population of size N, a strongly selected allele, which is destined to go to fixation, is introduced at time 0. If it is assumed that there is no dominance, the effect of the selected allele on existing heterozygosity at the neutral locus is then given by equation (19) of Stephan et al. (1992), as

$$H_1 = H_0 h, \tag{1a}$$

where

$$h = \frac{2c}{s}\,\alpha^{-2c/s}\Gamma\left(-\frac{2c}{s}, \frac{1}{\alpha}\right). \tag{1b}$$

H_0 is the expected heterozygosity at the neutral locus at time zero, when the selected allele occurs in the population, and H_1 is the expected heterozygosity at the time when the selected allele reaches fixation. s is the selective advantage of the favored allele per

generation, c is the rate of crossing-over between the neutral and selected locus, and $\alpha = 2Ns$. Γ denotes the incomplete gamma function.

In the report by Stephan et al. (1992), equation (1b) was derived by analyzing the diffusion process in a population directly, rather than a sample of genes. For further analysis, it is convenient to relate h to a sample quantity. Since nucleotide heterozygosity is given by the number of differences between two randomly sampled genes, it is sufficient to consider a sample of size 2. The genealogy of a sample is commonly described such that the time is running backward. Using this approach, Kaplan et al. (1989) have shown that the reduction in expected heterozygosity (in a population), h, that is due to a single hitchhiking event can be related to a sample quantity as follows [see the arguments leading up to equation (16) of Kaplan et al. (1989)]: Since the duration of the selective phase is very short, two genes sampled just after fixation of the selected allele can only be heterozygous [the probability of this event is H_1; see eq. (1a)] if they do not have a common ancestor during the selective phase and if their ancestral genes at time 0 (defined above) are different (the probability of this event is H_0). The latter two events are independent. Thus, it follows from equation (1b) that the reduction in expected heterozygosity h is equal to the probability of entering the selected phase with two ancestral genes and exiting it with two ancestral genes. In other words, h is equal to the probability that the neutral locus escapes hitchhiking by recombining away from the favored allele, while the favored allele is on its way to fixation. For a given value of s, this probability depends critically on the amount of recombination between the neutral and selected locus. If c is large, then the probability of escape is close to 1. On the other hand, if linkage is tight, then the neutral locus undergoes hitchhiking with high probability. In the case of complete linkage, h becomes 0 (see Stephan et al. 1992, table 1). Equations (1) are fundamental to the formulation of genetic hitchhiking as a steady-state process.

Steady-State Model of Recurring Substitutions

We assume that selected substitutions occur randomly along a chromosome according to a time-homogeneous Poisson process at a rate ν per nucleotide site (i.e., ν is the expected number of strongly selected substitutions per nucleotide site per generation). We consider the effect of these substitutions on polymorphism at a given neutral nucleotide site. Let the rate of crossing-over per nucleotide in the region, in which this neutral site is located, be ρ. Furthermore, let k_h be the expected number of selected substitutions (per $2N$ generations) that drag the neutral locus to fixation. The subscript indicates that this quantity depends on h, the probability that the neutral locus escapes a hitchhiking event, defined in equation (1b). If it is assumed that the chromosome is continuous, k_h can be calculated as follows: Consider selected substitutions that occur between m and $m + dm$ nucleotides away from the neutral site under study. Their expected number per generation is νdm. Furthermore, write c as a function of the recombination rate ρ per nucleotide and the distance m, i.e., $c = \rho m$, and write h as $h(\rho m)$. Then the rate at which a neutral polymorphism undergoes hitchhiking caused by selected substitutions between m and $m + dm$ nucleotides away is given by $\nu[1 - h(\rho m)]dm$. Since selected substitutions can occur on both sides of the neutral reference site, this expression has to be multiplied by a factor 2. To calculate k_h, we have to sum up these infinitesimal contributions over the entire chromosome. Scaling back to the original variable c and taking into account that k_h is measured in units of $2N$ generations, we find [see Stephan et al. 1992, eq. (20)]

$$k_h = 4N\frac{\nu}{\rho}\int_0^{M^*}[1 - h(c)]dc\,. \qquad (2)$$

Thus, k_h is inversely proportional to the recombination rate ρ in the region where the neutral site under consideration is located. The integral is independent of ρ. The integration goes over the entire neighborhood of the neutral locus up to a maximal recombinational distance, M^*, from the neutral site, which corresponds to a maximum physical distance of M^*/ρ nucleotides (Kaplan et al. 1989). The cutoff of the integration interval at M^* accounts for the assumption that, at any one time, at most only one substitution is on its way to fixation. We have to make this assumption for technical reasons, because we are considering a two-locus model. The consequences of this assumption have been examined in detail by Kaplan et al. (1989). We use their method to calculate M^* (see Appendix).

Now we ask what is the amount by which the level of neutral heterozygosity is reduced by recurring substitution events, assuming that the selected substitutions are sufficiently infrequent. The coalescent for a sample of two genes can be described as a simple stochastic process (jump process) with two states, consisting of two ancestral genes (state 2) or one ancestral gene (state 1). As the process goes back in time, it can change from state 2 to state 1 in two ways: either by the occurrence of a common ancestor or by the occurrence of a selected substitution that drags along the neutral locus to fixation. For two randomly sampled genes, the expected time to the most recent common ancestor is $2N$ generations. Therefore, on the time scale of $2N$ generations, the rate of occurrence of a coalescent event is 1, and the rate of occurrence of a hitchhiking event is k_h. Furthermore, the rate of occurrence of a coalescent or a hitchhiking event is $1 + k_h$, and the time back until one of these events occurs is $1/(1 + k_h)$. That means that the expected time of the coalescent, given by the sum of the two branch lengths, is $2/(1 + k_h)$. The expected time $E(T)$, measured in generations, is then given by $4N/(1 + k_h)$. To obtain the expected nucleotide heterozygosity we have to multiply the expected time of the coalescent by the neutral nucleotide mutation rate μ (Kaplan et al. 1989). Hence,

$$H = H_{\text{neu}}\frac{1}{1 + k_h}\,, \qquad (3)$$

where $H_{\text{neu}} = 4N\mu$ is the expected heterozygosity per nucleotide under neutrality (Kimura 1983, chap. 8.6). Equation (3) is equivalent to the first formula of equation (18) in Kaplan et al. (1989).

Equation (3) leads to an explicit functional relationship between average nucleotide heterozygosity and recombination rate ρ, as follows: Substituting u for $-2c/s$ in equation (1b) and M for $2NM^*$ in equation (2) gives

$$k_h = -\alpha\frac{\nu}{\rho}\int_0^{-2M/\alpha}\left[1 + u\alpha^u\Gamma\left(u, \frac{1}{\alpha}\right)\right]du\,. \qquad (4)$$

Then, denoting the integral on the right-hand side as $-I_M$ leads from equation (3) to

$$H = H_{\text{neu}} \frac{\rho}{\rho + \nu\alpha I_M}, \tag{5}$$

where $\rho > \nu/\Lambda_{\max}$, and $\Lambda_{\max}$ is a constant. The technical details regarding the domain of the right-hand side of this equation are given in the Appendix. Equation (5) does not hold for very low recombination rates, because we assume that, at any one time, at most one strongly selected mutation is going to fixation.

In the Appendix, we also establish the result that the right-hand side of equation (5) depends only on α, other than the obvious parameters ν and H_{neu}. Moreover, we demonstrate in figure 1 that αI_M is more or less linear over a wide range of α. Therefore, we obtain the important result that the reduction of heterozygosity below the neutral level as a function of the recombination rate ρ is essentially determined by a single parameter, $\alpha\nu$.

Transient Neutral Recovery Period

To describe the recovery of neutral polymorphism after a selective sweep under mutation pressure and drift, several approaches can be taken (Crow and Kimura 1970, chap. 8.5; Li 1977; Tajima 1989). For example, the expectation of the average number of nucleotide differences, S_2, among two sequences randomly sampled from a population is given by Tajima's (1989) equation (8), as

$$S_2(t) = \theta + [S_2(0) - \theta]\exp[-t/(2N)], \tag{6a}$$

where t is time measured in generations since the last selective sweep and $\theta = 4Nv$, where v is the neutral mutation rate per sequence. On a per-nucleotide basis, this equation reads

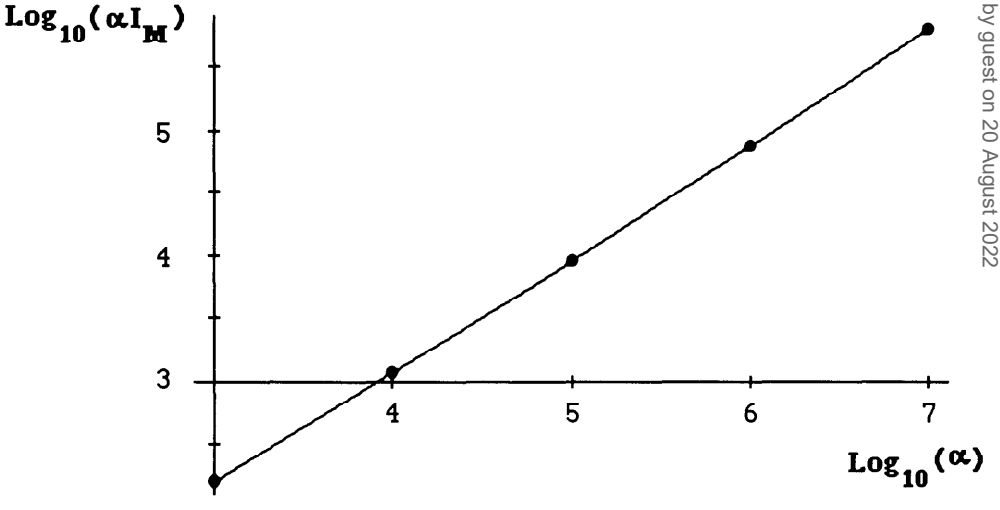

FIG. 1.—Rate of selected sweeps, which increases approximately linearly as a function of selection intensity. As eq. (4) shows, this rate is given by $-\nu/\rho(\alpha I_M)$, where $-I_M$ denotes the integral in eq. (4), ν is the frequency of selected mutations per base pair per generation, and ρ is the recombination rate per base pair per generation.

$$H(t) = H_{neu} + [H(0) - H_{neu}]\exp[-t/(2N)] . \quad (6b)$$

Equation (6b) requires knowledge of $H(0)$, the level of average nucleotide heterozygosity after a hitchhiking event took place. According to equations (1), this value is generally not known, unless heterozygosity before the selective sweep is known. An important exception is the case of 0 recombination. In regions of 0 recombination, hitchhiking reduces the level of nucleotide heterozygosity to 0, independently of the level of variation before the selective sweep. In this case, the recovery of average nucleotide heterozygosity can be described as

$$\frac{H(t)}{H_{neu}} \approx \frac{t}{2N}, \quad (6c)$$

as long as $t \ll 2N$, which we assume throughout this paper.

The case of 0 recombination is important for the following reasons: As mentioned above, we have to exclude chromosomal regions of very small recombination rates from the steady-state analysis, because we allow only one selected substitution at a time. However, a more important reason is that the recovery of neutral polymorphism changes the value of nucleotide heterozygosity existing immediately after a selective sweep most dramatically in regions of very low recombination rates. This can be demonstrated by using the simple model of the recovery phase shown in equation (6b). When $t \ll 2N$, this equation can be rewritten as

$$\frac{H(t) - H(0)}{H(0)} \approx \left(\frac{H_{neu}}{H(0)} - 1\right)\frac{t}{2N} . \quad (7)$$

The left-hand side denotes the relative departure of nucleotide heterozygosity, after a recovery period of t generations, from its initial value, i.e., the level of heterozygosity immediately after a hitchhiking event has occurred. The right-hand side shows that this deviation increases with increasing $H_{neu}/H(0)$. The ratio $H_{neu}/H(0)$ is expected to be larger in regions of low recombination rates. We illustrate this with two examples. Suppose that $H_{neu} = 0.008$ and that t, the time back to the last hitchhiking event, is $0.05N$. These numbers may be appropriate for *Drosophila melanogaster* (see next section). In the first example, we assume that at time t a nucleotide heterozygosity level of $H(t) = 0.004$ is observed in a population. Then, according to equation (6b), $H(0) = 0.0039$, and the relative deviation of $H(t)$ from $H(0)$, according to equation (7), is only 2.6%. In the second example, we assume that a value $H(t) = 0.0005$ is observed in a region of very low recombination. Then $H(0) = 0.00031$, and the relative deviation is 61.3%. Since $H(0)$ is expected to be smaller than the steady-state level for a given recombination value, the relative deviation of $H(t)$ from the steady-state level may be somewhat smaller. Nonetheless, these considerations suggest that the steady-state approach to hitchhiking is justified only for chromosomal regions with sufficiently high recombination rates, because, the relative deviation of the level of variation, at any point in time, from its steady-state value is then expected to be small. However, in regions of very low recombination, levels of variation may be rather different from their equilibrium values or their levels immediately after a hitchhiking event, as they depend more strongly on the time elapsed since the last sweep.

Time Scales

We consider here two asymptotic approaches to the hitchhiking process, one for regions of very low recombination rates and one for intermediate to high recombination rates. These are based on several assumptions about time scales. In the steady-state model, we assume that only one selected substitution is happening at a time. For the recovery period, we make the seemingly conflicting assumption that the time back to the last hitchhiking event is short—or, in other words, that the frequency of selected substitutions is high. As Kaplan et al. (1989) have emphasized, selective sweeps are virtually instantaneous, when time is measured in units of $2N$ generations. Therefore, our assumption of only one selective sweep at a time appears to apply to many systems, even if the frequency of selected substitutions is high. We illustrate this with the following example, which may be appropriate for *D. melanogaster*. Assume a moderate value of $\alpha = 10^4$. Then the expected time between the introduction and fixation of the favored allele is $\sim 0.0042N$ generations (see Kaplan et al. 1989, table 1). This is much shorter than the average time between the occurrence of successive selective substitutions, which we calculate as $\sim 0.14N$ generations for *D. melanogaster* (see next section).

An Application

In this section, we use the data compiled by Begun and Aquadro (1992) to estimate the parameter αv of the steady-state hitchhiking model. Begun and Aquadro presented for each of 20 gene regions in *Drosophila melanogaster* a value for average nucleotide heterozygosity (π) and a coefficient of exchange. We identify π with H of equation (5). The coefficient of exchange for a gene region was obtained by selecting genetically defined loci that flank the region of interest and by calculating the difference quotient of the distance, in map units, between the flanking loci and the number of polytene bands. To a first approximation, the coefficient of exchange should therefore be proportional to the recombination rate per nucleotide, ρ, in the gene region of interest. A proportionality constant can be calculated by assuming that the coefficient of exchange for *white* corresponds to the average rate of intragenic recombination in this gene region. The latter is $\sim 2 \times 10^{-8}$/bp (Judd 1987). Using this number, we find a proportionality factor of 1.43×10^{-7}, by which the coefficient of exchange has to be multiplied to obtain ρ. To make X-linked and autosomal gene regions comparable with respect to recombination rate and nucleotide diversity, the estimates of these quantities have to be transformed, as described in the legend to figure 2.

To estimate the parameter αv from Begun and Aquadro's (1992) data, we partition the 20 loci into two groups. This is suggested by the analysis in the previous section. Group I comprises the 17 loci from regions of intermediate to high recombination rates and low to high π values, and group II consists of loci in regions of unmeasurably low recombination rates and very low π values (*su(f)*, *y-ac-sc*, ci^D). This classification into two groups may not be appropriate for all loci. First, we examine this classification with respect to levels of heterozygosity. The metallothionein gene region (*Mtn*) has a low coefficient of exchange (0.0083) and also a low level of nucleotide diversity (0.001). However, the level of nucleotide polymorphism in *Mtn* is not significantly reduced relative to that of the *Adh* 5′ flanking region, which was used as a standard (Lange et al. 1990). Thus, *Mtn* does certainly not belong to the group of loci that exhibit extremely low levels of heterozygosity, such as ci^D and *su(f)*. Based on DNA sequencing and four-cutter restriction mapping, the following estimates of average nucleotide diversity

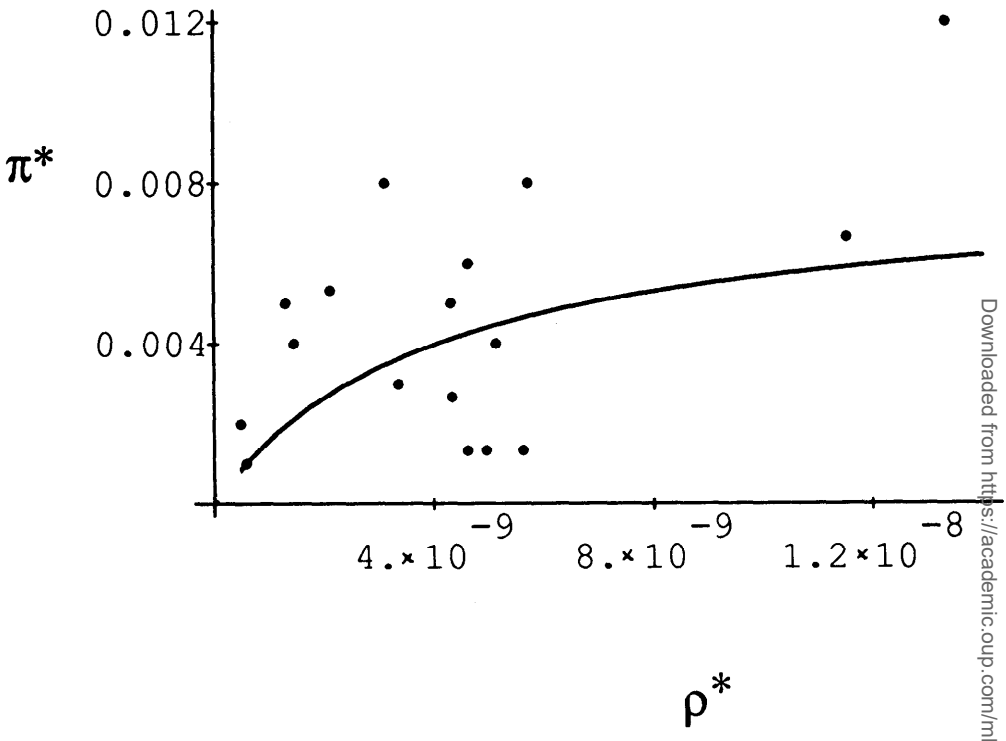

FIG. 2.—Nucleotide diversity vs. recombination rate. The data for 17 group I gene regions are from Begun and Aquadro (1992); the loci ci^D, *su(f)*, and *y-ac* from their table 1 are excluded, for reasons described in the text. The curve represents the best fit of eq. (5) to these data when the geometric mean method is used. ρ* = Transformed recombination coefficient. To make gene regions comparable, ρ values of X-linked loci are multiplied by $^2/_3$, and those of autosomal loci are multiplied by $^1/_2$. π* = Transformed nucleotide heterozygosity. π Values of X-linked loci are multiplied by $^4/_3$ (see Begun and Aquadro 1992).

have been found for the latter two group II loci: 0 for ci^D (Berry et al. 1991) and 0.0002 for *su(f)* (Langley et al. 1993). Another interesting case is the *y-ac-sc* region. While some analyses using six-cutter restriction enzymes showed that variation in this region is significantly reduced (Aguadé et al. 1989), other six-cutter studies did not find statistically significant reduction (Beech and Leigh Brown 1989; Eanes et al. 1989; Begun and Aquadro 1991). However, in a more recent study, Martín-Campos et al. (1992) reported statistical significance, on the basis of four-cutter analysis and a very large sample size of 245 lines. They found an average nucleotide diversity level of 0.00043 for *y-ac-sc*.

Second, we ask whether the above classification is justified with respect to recombination rate. The difference-quotient method for calculating the coefficient of exchange rests on a number of assumptions—that flanking loci exist, that the genetic map is sufficiently smooth around the region of interest, and that the polytene bands are spaced in a regular way and are clearly visible. For *su(f)*, which is located at the base of the X chromosome at the junction between β heterochromatin and euchromatin, the latter three assumptions are not met very well. The coefficient of exchange is likely to be lower than the value of 0.005 provided by Begun and Aquadro. For *y-ac-sc*, which is at the tip of the X chromosome, the first assumption is not met. If we apply the difference-quotient method to the very tip of the X chromosome and to *su*(w^a),

the first locus that has a genetic map position >0 and that is cytologically well defined at position 1E1-1E2 (Lindsley and Zimm 1992, p. 1118), we find a coefficient of exchange of 0.0033. This appears to be an overestimate for the coefficient of exchange in the *y-ac-sc* region, which is located at 1B1-1B4. It implies that Begun and Aquadro's value of 0.0045 for *y-ac* would also be too large. A lower value of the coefficient of exchange for *y-ac-sc* may be in better accordance with the observation of extensive linkage disequilibrium in this region, spanning distances of ≤80 kb (Macpherson et al. 1990; Martín-Campos et al. 1992).

For parameter estimation, equation (5) suggests a Lineweaver-Burk transformation of the group I data. The transformed model then becomes linear in $1/\rho$:

$$\frac{1}{\pi} = \frac{1}{H_{neu}} + \frac{1}{\rho}\frac{\nu\alpha I_M}{H_{neu}} = \beta_1 + \frac{1}{\rho}\beta_2 . \qquad (8)$$

The parameters β_1 and β_2 are determined from fitting equation (8) to the transformed data. Estimates of the parameters H_{neu} and $\nu\alpha I_M$ can be obtained from estimates of β_1 and β_2. According to equation (5), $H_{neu} = 4N\mu$ is the high recombination limit of average nucleotide heterozygosity. In the following estimation procedure, it is assumed that H_{neu} characterizes the whole genome of *D. melanogaster,* not individual gene regions, and that deviations from H_{neu} are exclusively due to hitchhiking.

For regression analysis, we used the geometric mean (GM) method, a model II regression procedure, because both variables are random (Sokal and Rohlf 1981, chap. 14.13). The measurements of π and the coefficient of exchange are subject to intrinsic measurement errors, as well as to stochastic fluctuations due to the evolutionary process. Therefore, it seems to be adequate to regress both variables on each other and to determine a mean value of both regression coefficients. For the GM method, this mean value is the GM. GM regression yields $\beta_1 = 125.54$ and $\beta_2 = 5.04 \times 10^{-7}$. The 95% confidence limits for β_2 are $L_1 = 2.51 \times 10^{-7}$ and $L_2 = 7.57 \times 10^{-7}$. From these estimates, we calculate $H_{neu} = 0.008$ and $\nu\alpha I_M = 4.03 \times 10^{-9}$. A value for $I_M = 0.075$ can be obtained from figure 1 by interpolation between $\alpha = 10^4$ and $\alpha = 10^6$. With these values, we estimate $\alpha\nu = 5.37 \times 10^{-8}$. Figure 2 shows the data and the fitted regression curve, when $H_{neu} = 0.008$ and $\nu\alpha I_M = 4.03 \times 10^{-9}$ are used as parameters. Finally, note that the condition for this analysis, $\rho > \nu/\Lambda_{max}$, is met for $\delta = 0.05$ but not for $\delta = 0.01$ (see table 1).

Confidence intervals of H_{neu} and $\alpha\nu$ are not provided by this method. However, it is possible to obtain a lower bound of $\alpha\nu$ by the following arguments. H_{neu} is likely to be larger than the level of nucleotide heterozygosity, H', obtained by averaging over all 20 gene regions, because these include also regions of reduced levels of heterozygosity. Therefore, we have $L_1 H' < L_1 H_{neu}$, and $L_1 H'$ is likely to be a lower bound of $\nu\alpha I_M$. Using the data set of Begun and Aquadro (1992) and correcting the heterozygosity values of the X-linked loci appropriately (see the legend to fig. 2), we find $H' = 0.004$. Therefore, $\alpha\nu$ is likely to be $> 1.34 \times 10^{-8}$.

To estimate this parameter more precisely, we need additional data. Data in high-recombination regions would help us to obtain a more reliable estimate of H_{neu}. This estimate could be then used to obtain also an upper bound for $\alpha\nu$. Furthermore, additional data in regions of low to intermediate coefficients of exchange ($\sim \geq 0.001$) would reduce the error in the observed π values, which is introduced by the recovery process, as described above. It would also be helpful to redo some of the measurements

of genetic variation in regions of intermediate to high recombination rates, to examine whether the large scatter of the data is due to sampling errors. Together, this additional work may narrow the confidence interval of αv considerably.

Finally, we make a crude calculation of the expected time back to the last hitchhiking event and of the expected duration of the recovery period. This calculation is to support the arguments about time scales given above. We use equation (6c) and the estimates of nucleotide diversity of the group II loci: 0 for ci^D (Berry et al. 1991), 0.0002 for $su(f)$ (Langley et al. 1993), and 0.00043 for *y-ac-sc* (Martín-Campos et al. 1992). The arithmetic mean of these loci, which takes into account the correction for X-linked and autosomal genes, is 0.00028. Therefore, when $H_{neu} = 0.008$ is used, the expected time back to the last selective sweep can be obtained from equation (6c), as $t = 0.07N$. Since the time between successive hitchhiking events is, on average, twice as long as the time back to the last hitchhiking event, we find that, on average, selective sweeps occur in regions of 0 recombination, such as the fourth chromosome, about every $0.14N$ generations. This result was obtained by considering only average nucleotide heterozygosity. To put confidence limits on this value, it is certainly worthwhile to develop more sophisticated techniques, as more data become available.

Discussion

The hitchhiking process can be considered as a series of two-step events, where a strongly selected substitution wiping out linked variation in a population alternates with a neutral recovery period in which polymorphism can build up via drift and mutation. The hitchhiking process has previously been modeled as a steady-state process with recurring selected substitutions, neutral mutation, random genetic drift, and recombination (Kaplan et al. 1989; Stephan et al. 1992). Here we analyze some properties of this model and demonstrate that the steady-state model describes the hitchhiking effect adequately, unless recombination rate is too low. The main result of our analysis of the steady-state model is that, given recombination rate, the reduction of neutral heterozygosity below the neutral level is characterized by a single parameter, αv.

This result implies that substitutions, satisfying αv = constant, have the same long-term effect on genetic variation. If the proposed hitchhiking model is adequate and this constant can be determined, some information on the upper tail of the frequency spectrum of selected substitutions can be obtained. Our analysis appears to imply that the frequency distribution of selection coefficients for strongly selected substitutions follows (perhaps only partially, up to a certain value of s) a power law of the form $v \propto s^{-1}$. This would mean that the frequency distribution of beneficial mutations is asymptotically of the form s^{-2}, since the probability of fixation for advantageous mutations is approximately $2s$. Such power laws are interesting, as they indicate a long-range behavior of a system, as opposed to the usual short-range exponential decay. Two fitness distributions have been commonly used in the literature. Both are for deleterious mutations, and both are of the exponential type: Ohta's (1977) purely exponential distribution of selection coefficients and Kimura's (1983, chap. 8.7) gamma distribution. They have been introduced mainly for theoretical reasons. It remains an open question to what extent they are confirmed empirically.

The steady-state hitchhiking model is used to analyze the data on DNA sequence variation in genomic regions of *D. melanogaster* compiled by Begun and Aquadro (1992). Our analysis suggests a partitioning of the loci into two groups: group I, which contains 17 loci from regions of intermediate to high recombination rates, and group

II, consisting of three genes located in regions of zero or very low recombination rates. Using the group I data, we estimate that the characteristic parameter of the steady-state model, αv, is likely to be $>1.3 \times 10^{-8}$. We conclude that more data are needed to determine the value of αv more precisely and hence the tail of the frequency spectrum of selected substitutions.

While the basic assumption of the steady-state hitchhiking model that at most only one selected substitution is on its way to fixation at a time does not seem to be a problem (at least not for *D. melanogaster*), our results suggest that this model has limitations in regions of very low recombination rates. However, this problem is not serious, when data from regions with sufficiently high recombination rates can be used to determine αv, as for *Drosophila.* It is more important at this point to consider alternative models, which predict a reduction of nucleotide diversity as a function of recombination rate. Such models may be used to explain some of the scatter in the data seen in *D. melanogaster,* especially in regions of intermediate recombination rates (see fig. 2). Furthermore, the estimate of αv may be influenced considerably if the reduction in nucleotide heterozygosity is partly due to other forces, such as deleterious mutations. Charlesworth et al. (submitted) have demonstrated that selection against deleterious alleles may have a similar effect on linked neutral polymorphism as directional selection. If hitchhiking with deleterious alleles plays a role, fewer selected sweeps, caused by advantageous mutations, and smaller selection coefficients are required to explain the observations.

Acknowledgments

We thank Brian Charlesworth, Jody Hey, and two anonymous reviewers for many valuable comments on this paper. T.H.E.W. was supported by a fellowship from the German Academic Exchange Service (DAAD).

APPENDIX

Properties of Equation (5)

We start from equation (4). For evaluating the integral on the right-hand side of equation (4), M needs to be defined. We used equation (20) of Kaplan et al. (1989) and determined $M = 2NM^*$ such that

$$\int_0^{M^*} [1 - h(c)]dc = \frac{1}{1+\delta} \int_0^{M_f^*} [1 - h(c)]dc \qquad \text{(A1)}$$

is satisfied. δ is a small positive constant, and $M_f = 2NM_f^*$ is chosen large (see below); the other parameters are as defined in the main text. The idea behind this procedure is to choose M as large as possible in order to take into account, around the neutral site, the entire neighborhood in which selected substitutions can occur and cause a hitchhiking effect. On the other hand, one has to restrict this region. Otherwise, the assumption that at most only one substitution is on its way to fixation might be violated. This choice seems to be arbitrary, but it guarantees that the integral on the left-hand side of equation (A1) will change only by a small amount δ when M^* is increased to M_f^*. Using equation (1b) for $h(c)$ and substituting u for $-2c/s$ and M for $2NM^*$, equation (A1) becomes

$$\int_0^{-2M/\alpha} \left[1 + u\alpha^u \Gamma\left(u, \frac{1}{\alpha}\right)\right] du = \frac{1}{1+\delta} \int_0^{-2M_f/\alpha} \left[1 + u\alpha^u \Gamma\left(u, \frac{1}{\alpha}\right)\right] du . \qquad \text{(A2)}$$

Table A1
Numerical Values of g^*, M, αI_M, and Λ_{max} for Different Values of α and δ

		M[b]		αI_M[c]		Λ_{max}[d]	
α	g^* [a]	$\delta = 0.01$	$\delta = 0.05$	$\delta = 0.01$	$\delta = 0.05$	$\delta = 0.01$	$\delta = 0.05$
10^3 ...	1.6×10^{-2}	301	235	160.5	154.4	0.00146	0.0050
10^4 ...	2.1×10^{-3}	2,396	1,756	1,164.0	1,119.7	0.00059	0.0028
10^5 ...	2.6×10^{-4}	18,696	13,594	9,103.2	8,756.4	0.00033	0.0024
10^6 ...	3.0×10^{-5}	152,350	110,845	74,806.3	71,956.5	0.00028	0.0024

[a] The time (in $2N$ generations) to fixation of a selected mutation, calculated by integrating formulas (5.51) and (5.52) of Ewens (1979).

[b] Determined as described in the text.

[c] $-I_M$ = the integral in eq. (4).

[d] Upper bound for ν/ρ; it is computed by using equation (22) of Kaplan et al. (1989).

M_f is chosen as the point where the integrand function in equation (A2), $g(u) = 1 + u\alpha^u\Gamma(u, 1/\alpha)$, is 1% away from its asymptotic limit, g^*, with respect to u (Kaplan et al. 1989). Because $h(c)$ is a good approximation (within 0.2%) of the exact function only as long as $2M/\alpha < 1$, we cannot use the limit of the integrands in equation (A2) as u approaches $-\infty$ [see the remarks after eq. (23b) in Stephan et al. 1992]. The exact asymptotic limit of g is equal to the fixation time of the selected mutant (in units of $2N$ generations). We calculated this limit, g^*, by using Ewens's (1979) diffusion approximation formulas (5.51) and (5.52). It follows from these formulas that g^* depends on α but not on N and s separately. Kaplan et al. (1989) showed that the diffusion approach is sufficiently accurate in this case. The results are listed in table A1, for δ values of 0.01 and 0.05. The assumption of at most one substitution event at any given time requires us also to specify a maximal value, Λ_{max}, for ν/ρ. We used equation (22) of Kaplan et al. (1989) to determine numerical values for Λ_{max} (table A1).

The above discussion reveals an important property of the steady-state model. Because the function defining M_f depends only on α, equation (A2) shows that M depends only on α as well, but not explicitly on N and s. It follows from equation (4) that k_h is only dependent on α, other than ν and ρ, and the same holds for H/H_{neu} in equation (5).

LITERATURE CITED

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

BEECH, R. N., and A. J. LEIGH BROWN. 1989. Insertion-deletion variation at the *yellow-achaete-scute* region in two natural populations of *Drosophila melanogaster.* Genet. Res. **53**:7–15.

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

———. 1992. Levels of naturally occurring DNA polymorphism are correlated 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.

CHARLESWORTH, B., M. T. MORGAN, and D. CHARLESWORTH. The effect of deleterious mutations on neutral molecular variation (submitted).

CROW, J. F., and M. KIMURA. 1970. An introduction to population genetics theory. Burgess, Minneapolis.

EANES, W. F., J. LABATE, and J. W. AJIOKA. 1989. Restriction-map variation in the *yellow-achaete-scute* region in five populations of *Drosophila melanogaster.* Mol. Biol. Evol. **6**:492–502.

EWENS, W. J. 1979. Mathematical population genetics. Springer, Berlin.

JUDD, B. H. 1987. The *white* locus in *Drosophila melanogaster.* Pp. 81–94 *in* W. HENNING, ed. Results and problems in cell differentiation. Vol. **14**: Structure and function of eukaryotic chromosomes. Springer, Berlin.

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

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

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

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

LI, W.-H. 1977. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics **85**:331–337.

LINDSLEY, D. L., and G. G. ZIMM. 1992. The genome of *Drosophila melanogaster.* Academic Press, San Diego.

MACPHERSON, J. N., B. S. WEIR, and A. J. LEIGH BROWN. 1990. Extensive linkage disequilibrium in the *achaete-scute* complex of *Drosophila melanogaster.* Genetics **126**:121–129.

MARTÍN-CAMPOS, J. M., J. M. COMERÓN, N. MIYASHITA, and M. AGUADÉ. 1992. Intraspecific and interspecific variation at the *yellow-achaete-scute* region of *Drosophila simulans* and *Drosophila melanogaster.* Genetics **130**:805–816.

MAYNARD SMITH, J., and J. HAIGH. 1974. The hitch-hiking effect of a favorable gene. Genet. Res. **23**:23–35.

OHTA, T. 1977. Extension of the neutral mutation drift hypothesis. Pp. 148–167 *in* M. KIMURA, ed. Molecular evolution and polymorphism. National Institute of Genetics, Mishima, Japan.

SOKAL, R. R., and F. J. ROHLF. 1981. Biometry. W. H. Freeman, San Francisco.

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

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

STEPHAN, W., T. H. E. WIEHE, and M. W. LENZ. 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41**:237–254.

TAJIMA, F. 1989. The effect of change in population size on DNA polymorphism. Genetics **123**:597–601.

BRIAN CHARLESWORTH, reviewing editor

Received November 20, 1992; revision received February 18, 1993

Accepted February 18, 1993