*Theoretical Population Biology* **53**, 272–283 (1998)
*Article No.* TP971346

# The Effect of Selective Sweeps on the Variance of the Allele Distribution of a Linked Multiallele Locus: Hitchhiking of Microsatellites

Thomas Wiehe*

*Department of Integrative Biology, University of California, Berkeley*

Received April 23, 1997

**Microsatellite variation and the mechanisms which are responsible for this variation have received much attention in the last few years. Most theoretical studies of microsatellite allele distributions, however, did not incorporate the evolutionary dynamics of linked sites. The dynamics is usually modeled by invoking a special mutation mechanism such as stepwise mutation, which leads to a stepwise increase or decrease of the number of motif repeats on the occasion of mutation. It is shown here that selection at a locus, which itself is not subject to mutation, but which is adjacent to a microsatellite locus has an influence on statistics of the microsatellite allele distribution, provided that mutation rates are low to intermediate, when compared to $1/t_1$, the inverse of the time to fixation of a linked favorable substitution. If mutation rates are high, as for example in humans, a selective effect upon the microsatellite locus, such as hitchhiking, will quickly be obscured by mutations. In particular, in the latter case, the model shows that no correlation is to be expected between recombination rates and variability of microsatellites—such as had been predicted and experimentally demonstrated for nucleotide variability and recombination rates in *Drosophila*. The presented model is a generalization of the two locus two allele hitchhiking model which had been studied by Stephan and co-workers.** © 1998 Academic Press

## 1. INTRODUCTION

The effect of selective sweeps on heterozygosity of a linked neutral locus has been studied by Kaplan *et al.* (1989) with the help of coalescence theory, by Ohta and Kimura (1975), and Stephan *et al.* (1992) using a diffusion approach. Previously, Maynard Smith and Haigh (1974) had addressed a similar question based on a deterministic model. Stephan *et al.* (1992) focused on the reduction in heterozygosity at a neutral two-allele locus triggered by the fixation of linked, selectively favored substitutions (hitchhiking). The reduction of heterozygosity below its neutral equilibrium value due to a selective sweep is described by a characteristic parameter: the ratio of twice the recombination rate between the selected and neutral sites and the selection coefficient, $2r/s$. The results of Stephan *et al.* (1992) rely on the assumption that the neutral locus, called $\mathscr{A}$, is biallelic. Also, they excluded mutation at the neutral locus from their model. In this article, I will generalize their model and allow for the neutral locus to be multiallelic. Furthermore, mutation among the (neutral) alleles is incorporated as a stepwise mutation model, as it is commonly used for microsatellites. In particular, I will derive an expression for the reduction of the variance of the allele distribution due to an adjacent selective substitution. A deterministic and a stochastic version, using the diffusion approach along the lines of Stephan *et al.* (1992), are developed.

* Current address: Thomas Wiehe, Abt. Genomanalyse, Institut für molekulare Biotechnologie, Postfach 100 813, D-07708 Jena, Germany. E-mail: twiehe@imb-jena.de.

0040-5809/98 $25.00

Variation at microsatellite loci, which are abundant in any eukaryotic species, has been the subject of numerous theoretical investigations (see Freimer and Slatkin, 1996, and references therein). Alleles at such loci are distinguished by differences in the number of repeats of simple sequence motifs. The mutation process is usually described by a random walk model where the state of the random walk corresponds to the number of motif repeats which are present in a specific allele. The symmetric random walk model with small stepsize has been found to misrepresent actual distributions in allele size in some cases (Di Rienzo *et al.*, 1994). To overcome this problem other mutation processes, like the two-phase mutation model (Di Rienzo *et al.*, 1994), have been proposed. Properties of microsatellite variability for symmetric mutation models with arbitrary step sizes have been derived by Zhivotovsky and Feldman (1995). It has been repeatedly suggested that some form of selection acting at simple sequence repeat (SSR) loci, like directional selection for longer repeats, has to be invoked in order to account for interspecific differences in average allele size (Ellegren *et al.*, 1995; Amos and Rubinstzein, 1996). Here, however, I assume neutrality for the microsatellite locus and concentrate on the interaction with a second, linked locus where selection is operating. I show that selective sweeps at the second locus can cause a reduction of the variance of the microsatellite allele distribution when compared to its equilibrium value. The strength of this effect depends not only on the mutation rate but also on the mutation mechanism: mutation rates may be constant across alleles or not. Selective sweeps generally lead to an excess of one allele. The excess will gradually fade away under the action of mutation. Using coalescence theory, Slatkin (1995) investigated hitchhiking of microsatellites before. However, he assumed that the selected site is completely linked to the microsatellite locus and that variation, after being temporarily wiped out, is reestablished after a selective sweep under the forces of drift and mutation only. Here, I will drop the assumption of complete linkage and show that the strength of the hitchhiking effect depends on a combination of the recombination and mutation rates and the selection coefficient.

## 2. THE MODEL

The ordinary differential equation (ODE) (Eq. (1)) describes the deterministic dynamics of a multiallele two-locus model, subject to the action of mutation, selection, and recombination. The $x_{ij}$ are haplotype frequencies. The first index ($i$) corresponds to allele $A_i$ at a neutral locus ($\mathscr{A}$) and the second index ($j$) to allele $B_j$ at a neighboring selected locus ($\mathscr{B}$):

$$\dot{x}_{ij} = (1-r)\, x_{ij} w_{ij} + r \sum_{m,n} x_{im} x_{nj} w_{im,nj} - x_{ij} \bar{w} + \sum_{m,n} m_{ij,mn} x_{mn}. \qquad (1)$$

Without recombination ($r=0$) this ODE is the diploid mutation-selection equation in its decoupled form (Akin, 1979): $w_{ij}$ are the marginal fitnesses of haplotypes ($w_{ij} = \sum_{m,n} x_{mn} w_{ij,mn}$); $m_{ij,mn}$ is the mutation rate with which an *mn*-haplotype becomes an *ij*-haplotype. The sum over the $x_{im} x_{nj} w_{im,nj}$ terms reflects the various possibilities to create *ij*-haplotypes from *im*- and *nj*-haplotypes through recombination. The haplotype frequencies are functions of time $t$, i.e., $x_{ij} = x_{ij}(t)$. Throughout, a dot [ $\dot{}$ ] indicates differentiation with respect to time. A finite-number-of-alleles model is assumed: alleles at locus $\mathscr{A}$ are numbered from 0 to $v$, representing the number of motif repeats. Locus $\mathscr{B}$ is biallelic ($b$ and $B$). Selection at $\mathscr{B}$ is directional and according to the scheme

| BB | Bb | bb |
|---|---|---|
| $1+2s$ | $1+s$ | $1$ |

with selection coefficient $s$. Since $\mathscr{A}$ is neutral the fitness parameters $w_{i.,j.}$ are independent of $i$ and $j$:

$$w_{iB,jB} = 1+2s,$$
$$w_{iB,jb} = w_{ib,jB} = 1+s,$$
$$w_{ib,jb} = 1$$

for all $i$, $j$. Let allele $B$ be introduced into the population at time $t_0 = 0$ with frequency $x_B(t_0) = \varepsilon$. It is linked to one of the alleles $A_i$ at locus $\mathscr{A}$. It replaces the wildtype $b$ as it sweeps through the population. The substitution process takes

$$t_1 = \frac{-2}{s} \log(\varepsilon)$$

generations.

With the stepwise mutation model for the $\mathscr{A}$-locus and disregarding mutation at $\mathscr{B}$, the mutation rates are

$$m_{iB,jb} = m_{ib,jB} = 0,$$
$$m_{iB,iB} = m_{ib,ib} = -(\mu_1(i) + \mu_2(i))$$

for all $i$, $j$,

$$m_{iB,\,(i+1)\,B} = m_{ib,\,(i+1)\,b} = \mu_2(i+1),$$
$$m_{(i+1)\,B,\,iB} = m_{(i+1)\,b,\,ib} = \mu_1(i)$$

for all $i < v$, and

$$m_{ib,\,jb} = m_{iB,\,jB} = 0$$

if $|i-j| > 1$. Given allele $A_i$ of length $i$, then $\mu_1(i)$ is its upward and $\mu_2(i)$ its downward mutation rate. Different assumptions about the form of the function $\mu.(.)$ have been made: for instance, Valdes *et al.* (1993) and Bell (1996) (his "RWM" model) use a constant function $\mu.(.) = \mu$; i.e., the mutation rate is constant across alleles and alleles increase or decrease in size with the same probability (unbiased model). In other words, $\mu = \frac{1}{2}\hat{\mu}$, where $\hat{\mu}$ is the mutation rate of the locus. Slatkin (1995) (see also Garza *et al.*, 1995; Bell, 1996) suggests a linear function, where the mutation rate depends on the current allelic state such that short alleles tend to increase in size and long ones tend to decrease (biased model). This reflects tight regulation of copy number and the fact that experimentally established allele distributions are often centered around a common allele with an intermediate number of motif repeats. Inching towards generality, I will consider both mutation mechanisms below. For the latter model let

$$\begin{aligned}\mu_1(i) &= \mu(v-i),\\ \mu_2(i) &= \mu i,\end{aligned} \tag{2}$$

with a constant $\mu$. This definition differs from that by Slatkin (1995) by a scaling factor of $v$. If one defines $\mu = (1/v)\,\hat{\mu}$ and interprets $\hat{\mu}$ as the mutation rate of the locus, then allele $A_i$ will increase in size with probability $1-(i/v)$ and decrease in size with probability $(i/v)$, when a mutation occurs. The deterministic equilibrium distribution under the mutation scheme (2) is binomial with parameters $v$ and 1/2.

Represented in a square matrix the mutation terms form two types of submatrices

$$M = \begin{pmatrix} \overbrace{M_1}^{b} & \overbrace{M_2}^{B} \\ M_2 & M_1 \end{pmatrix} \begin{matrix} \}\,b \\ \}\,B \end{matrix}.$$

Submatrix $M_1$ contains those entries which refer to mutation at locus $\mathscr{A}$ only. The allele at $\mathscr{B}$ is not altered. Entries in submatrix $M_2$ refer to changes at locus $\mathscr{B}$ (which are excluded here). For instance, for $v = 2$ and under the assumptions made, the matrix is

$$M = \begin{pmatrix} -\mu_1(0) & \mu_2(1) & 0 & 0 & 0 & 0 \\ \mu_1(0) & -\mu_1(1)-\mu_2(1) & \mu_2(2) & 0 & 0 & 0 \\ 0 & \mu_1(1) & -\mu_2(2) & 0 & 0 & 0 \\ 0 & 0 & 0 & -\mu_1(0) & \mu_2(1) & 0 \\ 0 & 0 & 0 & \mu_1(0) & -\mu_1(1)-\mu_2(1) & \mu_2(2) \\ 0 & 0 & 0 & 0 & \mu_1(1) & -\mu_2(2) \end{pmatrix}.$$

The effect of a selective sweep on the allele distribution at the linked neutral locus is here described in terms of the average reduction in the variance of the distribution of the number of SSRs. In the special case $v = 1$ (i.e., two neutral alleles) this quantity coincides with average reduction in heterozygosity. The term *average* means that the reduction is calculated for all possible initial conditions (when $B$ arises at $t_0$ it may be linked to any of the alleles $A_i$) and is then weighted according to the frequencies of alleles $A_i$ at $t_0$. For this purpose it is convenient to transform the haplotype frequencies, as given in Eq. (1), into conditional frequencies (cf. Maynard Smith and Haigh, 1974; Ohta and Kimura, 1975). Let $y_{i\,|\,B}$ be the frequency of allele $A_i$ at $\mathscr{A}$, conditioned on that it is linked to allele $B$ (to $b$, resp.) at locus $\mathscr{B}$. Then, Eq. (1) becomes

$$\dot{y}_{i\,|\,b} = r x_B(y_{i\,|\,B} - y_{i\,|\,b}) + [M \cdot y_b]_i, \qquad 0 \leqslant i \leqslant v, \tag{3}$$

$$\dot{y}_{i\,|\,B} = r(1-x_B)(y_{i\,|\,b} - y_{i\,|\,B}) + [M \cdot y_B]_i, \qquad 0 \leqslant i \leqslant v, \tag{4}$$

$$\dot{x}_B = s x_B(1-x_B). \tag{5}$$

The last terms in Eqs. (3) and (4) contain the contributions due to mutation. The vectors $y_b$ and $y_B$ are

$$y_b = (y_{0\,|\,b}, ..., y_{v\,|\,b}, 0, ..., 0)^T,$$
$$y_B = (0, ..., 0, y_{0\,|\,B}, ..., y_{v\,|\,B})^T$$

($T$ means transposition). A particular advantage of the transformation is that explicit selection terms are absent from Eqs. (3) and (4). The frequency of allele $A_i$ in the new variables is $(1-x_B)\, y_{i\,|\,b} + x_B y_{i\,|\,B}$ and therefore the average repeat number, $E(t)$, is

$$E(t) = \sum_{i=0}^{v} i((1-x_B(t))\, y_{i\,|\,b}(t) + x_B(t)\, y_{i\,|\,B}(t)).$$

Similarly, the variance of repeat number, $V(t)$, is

$$V(t) = \sum_{i=0}^{v} (i - E(t))^2\, ((1-x_B(t))\, y_{i\,|\,b}(t) + x_B(t)\, y_{i\,|\,B}(t)).$$

Note that, at $t = t_1$, the term $(1-x_B(t))\, y_{i\,|\,b}(t)$ is negligible, since $1 - x_B(t_1) = \varepsilon$.

The variance is invariant with respect to a translation of the interval $[0, v]$ to $[0+c, v+c]$ for any constant $c$. Therefore, the absolute repeat numbers are unimportant for the purpose here. What matters is only the difference between minimal and maximal copy numbers.

## 3. RESULTS

### 3.1. No Mutation

First, let us consider the case $M = 0$. As is obvious from Eqs. (3) and (4), the ODE system is only partially coupled; there are only pairs of coupled equations $y_{i\,|\,b}$ and $y_{i\,|\,B}$, for each $i$. All equations also contain $x_B$. However, the ODE for $x_B$ can be solved directly and the solution can then be inserted into Eqs. (3) and (4):

$$x_B(t) = \frac{\varepsilon}{\varepsilon + (1-\varepsilon)\, e^{-st}}. \tag{6}$$

The dynamics of any pair of coupled equations is identical to the one of the two allele model. Therefore, $V$ is, up to a factor of two, identical with the heterozygosity. Using the arguments by Stephan *et al.* (1992), one can calculate the variance in copy number after the selective sweep, $V(t_1)$, relative to the variance before the selective sweep, $V(t_0)$. In particular, making use of the fact that the reduction in heterozygosity is independent of the initial frequencies of the neutral alleles, one obtains

$$\frac{V(t_1)}{V(t_0)} = 1 - \varepsilon^{2r/s}. \tag{7}$$

Note, however, that averages here are taken over $v+1$, instead of two, possible initial combinations of the selected allele $B$ with any of the alleles $A_i$. At $t_0$, any combination $A_i/B$ is realized with probability $x_{ib}(t_0)$. It turns out that the right side of Eq. (7) is a lower bound to the case with mutation.

### 3.2. Mutation

When mutation terms are present Eqs. (3) and (4) no longer split into pairs of coupled equations. Analytical solutions (except for the special case $v = 1$; see the Appendix) are much harder to obtain. The following results have been derived by a partially numerical, partially analytical approach.

Numerical integration of system Eqs. (3) and (4) for a large variety of coefficients suggests that the ratio $V(t_1)/V(t_0)$ can be very well described by introducing into Eq. (7) a single additional parameter $\alpha$, which is independent of the recombination rate if $r$ is small enough (roughly, $r < s/2$; cf. Table I):

$$\frac{V(t_1)}{V(t_0)} = 1 - \alpha\varepsilon^{2r/s}. \tag{8}$$

For mutation rates as in Eq. (2) $\alpha$ can be determined analytically. It will be shown that, under this assumption, $\alpha$ does not explicitly depend on $v$. This, together with the fact that $\alpha$ is independent of $r$ for $v = 1$ (see Appendix), proves that $\alpha$ is indeed independent of $r$ for any $v$. To determine $\alpha$, consider the zero-recombination limit first. In this case, Eqs. (3) and (4) are homogeneous linear and

**TABLE 1**

**Independence of α and *r***

| $r$ | Biased mutation rates | | | Unbiased mutation rates | | |
|---|---|---|---|---|---|---|
| | $n^a = 2$ | $n = 10$ | $n = 20$ | $n = 2$ | $n = 10$ | $n = 20$ |
| $10^{-8}$ | 0.4528 | 0.4528 | 0.4528 | 0.4528 | 0.9584 | 0.9888 |
| $10^{-7}$ | 0.4528 | 0.4528 | 0.4528 | 0.4528 | 0.9584 | 0.9888 |
| $10^{-6}$ | 0.4528 | 0.4528 | 0.4528 | 0.4528 | 0.9584 | 0.9888 |
| $10^{-5}$ | 0.4528 | 0.4528 | 0.4528 | 0.4528 | 0.9584 | 0.9888 |
| $10^{-4}$ | 0.4529 | 0.4529 | 0.4530 | 0.4529 | 0.9587 | 0.9891 |
| $10^{-3}$ | 0.4679 | 0.4679 | 0.4693 | 0.4679 | 0.9904 | $>1$ |
| $10^{-2}$ | $>1$ | $>1$ | $>1$ | $>1$ | $>1$ | $>1$ |

*Note.* Values for $\alpha$ are calculated based on numerical integration (Runge–Kutta method from Press *et al.* (1992)) of system (3) to (5) from $t_0$ to $t_1 = -2/s \log(\varepsilon)$ (fixation time of a single selective substitution). $\alpha$ is independent of $n$, the number of neutral alleles, for the biased mutation model, but not so for the unbiased model. Parameters: $\varepsilon = 1/(2N)$; $N = 10^4$; $\mu = 10^{-4}$; $s = 10^{-2}$.

[a] $n = v + 1$.

independent of $x_B$. The eigenvalues $\lambda_k$, $0 \leqslant k \leqslant v$ of the matrix $M_1$ are

$$\{-2\mu v, -2\mu(v-1), ..., -2\mu, 0\}. \tag{9}$$

The eigenvalues are real and all distinct if $\mu \neq 0$. Thus, $M_1$ can be diagonalized by means of an invertible matrix $Q$, the inverse of which contains the (right) eigenvectors of $M_1$ as its columns. Furthermore, it can be shown that $Q$ may be chosen such that

$$Q^{-1} = 2^v Q.$$

With the coordinate transformation

$$y = Qx$$

($y$ here is distinct from the conditional frequencies $y_{.1.}$ in Sections 2 and 3.3), the task to solve the system

$$\dot{x} = M_1 x$$

can be reduced to solving the entirely decoupled system

$$\dot{y} = QM_1Q^{-1}y.$$

The solution is

$$y_j(t) = e^{\lambda_j t} y(t_0)$$

and, evaluated at $t_1$, one obtains

$$y_j(t_1) = \varepsilon^{-2\lambda_j/s} y_j(0), \tag{10}$$

where

$$y(0) = Qx(0).$$

If allele $B$ happened to be linked to allele $A_i$ at time $t_0$, then the initial condition in the original system is

$$x(t_0) = (0, ..., 0, 1, 0, ..., 0) = e_i,$$

with entry 1 at the $i$th position in vector $e_i$. Therefore, and because of the properties of $Q$, the initial condition of the transformed system is

$$y(t_0) = Qx(t_0) = Qe_i = 2^{-v}Q^{-1}e_i = 2^{-v}q_{.i},$$

where $q_{.i}$ is the $i$th (right) eigenvector of $M_1$ (= the $i$th column of $Q^{-1}$). In this representation, the first- and second-order moments of the $\mathscr{A}$-allele distribution can be calculated more easily. The entries $q_{ij} = q_{ij}^{(v)}$, $0 \leqslant i, j \leqslant v$, of the $(v+1) \times (v+1)$-matrix $Q^{-1} = (Q^{(v)})^{-1}$ can be determined recursively. In fact,

$$q_{ij}^{(v)} = q_{i-1, j-1}^{(v-1)} + q_{i, j-1}^{(v-1)},$$

if $0 < i$, $j$ and $i < v$,

$$q_{0j}^{(v)} = q_{0, j-1}^{(v-1)},$$
$$q_{vj}^{(v)} = q_{v-1, j-1}^{(v-1)}$$

if $0 < j$ and

$$q_{i0}^{(v)} = (-1)^{v-i}\binom{v}{i}$$

for $0 \leqslant i \leqslant v$. In particular, one has

$$\begin{aligned} q_{v-1, i} &= 2i - v, \\ q_{v, i} &= 1 \end{aligned} \tag{11}$$

for $0 \leqslant i \leqslant v$.

Now, the first-order moment, $E(t) = E(x(t))$, may be written as

$$E(x(t)) = \sum_{i=0}^{v} i x_i(t) = \sum_{i=0}^{v} i \sum_{j=0}^{v} q_{ij} y_j(t) = \sum_{j=0}^{v} y_j(t) \sum_{i=0}^{v} i q_{ij}.$$

As can be shown by induction over $v$,

$$\sum_{i=0}^{v} i q_{ij} = \begin{cases} 0, & j < v-1, \\ 2^{v-1}, & j = v-1, \\ v2^{v-1}, & j = v. \end{cases} \tag{12}$$

Thus,

$$E(x(t)) = 2^{v-1}(y_{v-1}(t) + v y_v(t)). \tag{13}$$

With the initial condition $x(t_0) = e_i$ and using the results from Eqs. (10) and (11), one finds

$$\begin{aligned} E_i(x(t_1)) &= 2^{v-1}(\varepsilon^{-2(-2\mu)/s} 2^{-v} q_{v-1, i} + v 2^{-v} q_{v, i}) \\ &= \left(i - \frac{v}{2}\right)\varepsilon^{4\mu/s} + \frac{v}{2}, \end{aligned}$$

where the index in $E_i$ refers to the initial condition $x(t_0) = e_i$, $0 \leqslant i \leqslant v$. Similarly to Eq. (12), one derives for the second-order moment

$$\sum_{i=0}^{v} i^2 q_{ij} = \begin{cases} 0, & j < v-2, \\ 2^{v-1}, & j = v-2, \\ v2^{v-1}, & j = v-1, \\ (v+1)\, v2^{v-2}, & j = v, \end{cases}$$

and, after some simplifications,

$$E_i^2(x(t_1)) = \frac{1}{2}\left(\left(\binom{v}{2} - 2i(v-i)\right)\varepsilon^{8\mu/s} + v(2i-v)\,\varepsilon^{4\mu/s} + \frac{v}{2}(v+1)\right).$$

The variance, $V_i(x(t_1))$, therefore is

$$V_i(x(t_1)) = E_i^2(x(t_1)) - (E_i(x(t_1)))^2 = \frac{v}{4}(1 - \varepsilon^{8\mu/s}).$$

Remarkably, $V_i$ is independent of $i$, the initial condition. With $V(t_0) = v/4$, which is the variance of the binomial equilibrium distribution before the selective sweep, one obtains the desired ratio in the no-recombination case:

$$V(t_1)/V(t_0) = 1 - \varepsilon^{8\mu/s}.$$

Comparing this to Eq. (8), it follows that

$$\alpha = \varepsilon^{8\mu/s}, \tag{14}$$

or, in terms of $t_1$, $\alpha = \exp(-4t_1\mu)$ and, in terms of the per locus mutation rate, $\alpha = \exp(-4t_1\hat{\mu}/v)$. The combined effect on $V$, which takes recombination and mutation into account, therefore, is

$$V(t_1)/V(t_0) = 1 - \varepsilon^{(8\mu + 2r)/s}. \tag{15}$$

In Eq. (15) $\alpha$ does not explicitly depend on $v$, the number of alleles at locus $\mathscr{A}$. This is not true for the mutation model with unbiased rates. In this case, although the eigenvalues may be computed analytically (see also Feldman *et al.*, 1997), the eigenvectors of $M_1$ are much more complicated and therefore, the allele frequency distribution at $t_1$, $x(t_1)$, is not easily accessible. I calculated $\alpha$ numerically for some parameter choices and compared it to formula (14) in Table II.

**TABLE 2**

**Effect of a Single Selective Sweep**

| Parameters | | | $\alpha$ | | | | | $r^*(\text{red} > 10\,\%)$[b] | |
|---|---|---|---|---|---|---|---|---|---|
| $N = 1/2\varepsilon$ | $\mu$ | $n$[a] | Biased | | | Unbiased | | Biased[c] | Unbiased[d] |
| | | | Eq. (14) | Eq. (23) | Simulation | RKI[d] | Simulation | | |
| $10^4$ | $10^{-3}$ | 2 | 0.0004 | 0.0032 | 0.0162 (0.0374) | 0.0004 | 0.0136 (0.0356) | ∅ | ∅ |
| | | 10 | 0.0004 | 0.0032 | 0.0067 (0.1453) | 0.6703 | 0.7292 (0.1535) | ∅ | 9.78 |
| | | 50 | 0.0004 | 0.0032 | 0.0034 (0.1575) | 0.9822 | 0.8734 (0.1238) | ∅ | 11.79 |
| | $10^{-4}$ | 2 | 0.4528 | 0.5187 | 0.6020 (0.2429) | 0.4528 | 0.5796 (0.2644) | 7.63 | 7.73 |
| | | 10 | 0.4528 | 0.5187 | 0.5816 (0.1852) | 0.9584 | 0.8581 (0.1584) | 7.63 | 11.66 |
| | | 50 | 0.4528 | 0.5187 | 0.5737 (0.1734) | 0.9981 | 0.8946 (0.1076) | 7.63 | 11.87 |
| | $10^{-5}$ | 2 | 0.9238 | 0.8767 | 0.8930[#] | 0.9237 | 0.8884[#] | 11.23 | 11.47 |
| | | 10 | 0.9238 | 0.8767 | 0.8560 (0.1622) | 0.9956 | 0.8946[#] | 11.23 | 11.84 |
| | | 20 | 0.9238 | 0.8767 | 0.8544 (0.1265) | 0.9988 | 0.8961[#] | 11.23 | 11.86 |
| $10^6$ | $10^{-5}$ | 2 | 0.8904 | 0.9080 | 0.9190 (0.0718) | 0.8904 | 0.9199 (0.0764) | 7.54 | 7.62 |
| | | 10 | 0.8904 | 0.9080 | 0.9195 (0.0250) | 0.9937 | 0.9939 (0.0051) | 7.54 | 8.00 |
| | | 20 | 0.8904 | 0.9080 | 0.9178 (0.0213) | 0.9977 | 0.9975 (0.0028) | 7.54 | 8.01 |
| | $10^{-6}$ | 2 | 0.9885 | 0.9893 | 0.9908 (0.0183) | 0.9885 | 0.9919 (0.0110) | 7.90 | 7.97 |
| | | 10 | 0.9885 | 0.9893 | 0.9917 (0.0082) | 0.9994 | 0.9987 (0.0058) | 7.90 | 8.01 |
| | | 20 | 0.9885 | 0.9893 | 0.9911 (0.0064) | 0.9998 | 0.9995 (0.0014) | 7.90 | 8.01 |

*Note.* Simulation results are averages (SD) based on 1000 replicates. Selection coefficient in all cases $s = 0.01$. For $v > 1$ the neutral equilibrium allele distribution (which gives $V_0$) is obtained by simulation. For the two-allele case (i.e., $v = 1$) the theoretical equilibrium distribution (a Beta distribution with parameters $N$ and $\mu$) may be used instead. ∅: even with complete linkage the reduction is less than 10%. [#]: these values are upper bounds (some of the 1000 replicates produced an increase of $V$ compared to $V_0$; those are disregarded).

[a] Number of alleles ($= v + 1$) at locus $\mathscr{A}$.

[b] Maximal recombination rate such that $V_1/V_0 < 0.9$; all values have to be multiplied by $10^{-4}$.

[c] Analytical values, obtained by taking the inverse of Eq. (15).

[d] Numerical integration using the Runge–Kutta method with variable stepsize (Press *et al.*, 1992).

### 3.3. Finite Population Size

Relying on a diffusion approach, Stephan *et al.* (1992) derived an analytical analogue to Eq. (7) which incorporates the effects of random drift due to finite population size. Their formula (their Eq. (19)) for reduction in heterozygosity $H$ due to a selective sweep and for the two allele case without mutation is

$$\frac{H(t_1)}{H(t_0)} = \frac{2r}{s}\,(2Ns)^{-2r/s}\,\Gamma\left(-\frac{2r}{s}, \frac{1}{2Ns}\right). \qquad (16)$$

In Eq. (16), $N$ is the diploid population size and $\Gamma$ denotes the incomplete Gamma function.

For the mutation scheme as in Eq. (2), one can derive a formula which describes the reduction of $V$ in the multiallele case from Eq. (16) by introducing again an additional parameter. Guided by the result before, one may choose the form

$$\frac{V(t_1)}{V(t_0)} = 1 - \alpha\left(1 - \frac{H(t_1)}{H(t_0)}\right). \qquad (17)$$

The fact that $\alpha$ in Eq. (14) is independent of $v$ suggests that the dynamics during the selective phase may be approximated by a two-locus two-allele (instead of $v+1$ alleles) diffusion model as introduced by Ohta and Kimura (1975) and adapted by Stephan *et al.* (1992). The diffusion equation, a partial differential equation of the form $(\partial/\partial t) = \mathscr{L}$ ($\mathscr{L}$ is the differential operator), may be integrated over suitable functions $f$. In particular, (ordinary) differential equations for arbitrary moments of the expected allele frequencies can be derived when $f$ is successively replaced by $E(Y_{1|B})$, $E(1 - Y_{1|B})$, $E(Y^2_{1|B})$ and so forth. The differential equations for the first and second moments, and including mutation terms, are then

$$\begin{aligned}
\frac{dE(Y_{1|B})}{dt} &= r(1-x_B)\,E(Y_{1|b} - Y_{1|B}) \\
&\quad + E(\mu(1 - Y_{1|B}) - \mu Y_{1|B}), \\
\frac{dE(Y_{1|b})}{dt} &= rx_B E(Y_{1|B} - Y_{1|b}) \qquad (18) \\
&\quad + E(\mu(1 - Y_{1|b}) - \mu Y_{1|b}), \\
\frac{dx_B}{dt} &= sx_B(1 - x_B),
\end{aligned}$$

and

$$\begin{aligned}
&\frac{dE(Y^2_{1|B})}{dt} \\
&\quad = E\left(\frac{Y_{1|B}(1 - Y_{1|B})}{2Nx_B} + 2r(1-x_B)\,Y_{1|B}(Y_{1|b} - Y_{1|B})\right. \\
&\qquad \left. + 2\mu Y_{1|B}((1 - Y_{1|B}) - Y_{1|B})\right), \\
&\frac{dE(Y_{1|B}Y_{1|b})}{dt} \\
&\quad = E(r(1-x_B)\,Y_{1|b}(Y_{1|b} - Y_{1|B}) \qquad (19) \\
&\qquad + \mu Y_{1|b}((1 - Y_{1|B}) - Y_{1|B}) \\
&\qquad + rx_B Y_{1|B}(Y_{1|B} - Y_{1|b}) \\
&\qquad + \mu Y_{1|B}((1 - Y_{1|b}) - Y_{1|b})), \\
&\frac{dE(Y^2_{1|b})}{dt} \\
&\quad = E\left(\frac{Y_{1|b}(1 - Y_{1|b})}{2N(1-x_B)} + 2rx_B Y_{1|b}(Y_{1|B} - Y_{1|b})\right. \\
&\qquad \left. + 2\mu Y_{1|b}((1 - Y_{1|b}) - Y_{1|b})\right).
\end{aligned}$$

The random variables $Y_{1|.}$ are the conditional relative frequencies of allele $A_1$ at locus $\mathscr{A}$, conditioned on that they are linked to either allele $b$ or $B$ at locus $\mathscr{B}$ (for the two allele case $Y_{0|.} = 1 - Y_{1|.}$). $E$ is the expectation of the random variables with respect to a transition density function $\phi$, which fortunately does not need to be specified in detail; $x_B$ is the (unconditional) frequency of allele $B$. Since selection is assumed to be strong it is treated as a deterministic quantity and, together with the initial condition $x_B(t_0) = \varepsilon = 1/(2N)$, is entirely determined by Eq. (6).

In the zero-recombination limit, system (18) and (19) becomes inhomogeneous linear. In fact, in order to determine the variance at locus $\mathscr{A}$ at time $t_1$, one needs, since $B$ will be fixed at $t_1$, only to consider the two equations

$$\begin{aligned}
\frac{dE(Y_{1|B})}{dt} &= \mu E(1 - 2Y_{1|B}), \\
\frac{dE(Y^2_{1|B})}{dt} &= E\left(\frac{Y_{1|B}(1 - Y_{1|B})}{2Nx_B} + 2\mu Y_{1|B}(1 - 2Y_{1|B})\right),
\end{aligned}$$

subject to the two possible initial conditions

$$E(Y_{1|B})(t_0) = 0, \qquad E(Y^2_{1|B})(t_0) = 0,$$

or

$$E(Y_{1\,|\,B})(t_0)=1, \qquad E(Y_{1\,|\,B}^2)(t_0)=1.$$

To calculate the variance it suffices to solve for the difference

$$z=E(Y_{1\,|\,B})-E(Y_{1\,|\,B}^2).$$

The differential equation for $z$ is

$$\dot{z}=\mu-\left(\frac{1}{2Nx_B}+4\mu\right)z, \tag{20}$$

subject to the (only one possible) initial condition $z(t_0)=0$. The solution in integral form is

$$z(t)=\mu\,\frac{\int_0^t \exp(\int_0^\tau 4\mu+(\varepsilon/x_B(\eta))\,d\eta)\,d\tau}{\exp(\int_0^t 4\mu+(\varepsilon/x_B(\tau))\,d\tau)}, \tag{21}$$

with $\varepsilon=1/(2N)$. Defining $f(t)=s^{-1}e^{-st}-t(4\mu+\varepsilon)$, Eq. (21) somewhat simplifies to

$$z(t)=\mu e^{f(t)}\left(\int_0^t e^{-f(\tau)}\,d\tau\right).$$

The initial variance $V(t_0)$ can be determined using the known density function of the one locus two allele model with mutation (cf. Crow and Kimura, 1970, p. 391), which is the density of a beta-distribution. Before the selective phase (up to $t_0$) only allele $b$ is present at $\mathscr{B}$. Therefore, one has a one-locus scenario. The equilibrium density is

$$\phi(p)=\frac{\Gamma(8\mu N)}{(\Gamma(4\mu N)^2)}\,p^{4\mu N-1}(1-p)^{4\mu N-1},$$

and $p$ $(0\leqslant p\leqslant 1)$ is the frequency of one of the neutral alleles, $A_1$, say. The variance is $V(t_0)=E(A_1)-E(A_1^2)$, with the moments taken with respect to $\phi$. Therefore,

$$V(t_0)=\frac{2N\mu}{1+8N\mu}. \tag{22}$$

Finally, if $r=0$, $(H(t_1)/H(t_0))=0$ and Eq. (17) turns into

$$\frac{V(t_1)}{V(t_0)}=1-\alpha.$$

Therefore,

$$\alpha=1-\frac{z(t_1)}{V(t_0)}. \tag{23}$$

Numerical values of $\alpha$, calculated according to Eqs. (14) and (23) for various parameters are listed in Table II. To check the validity for the multiallele case, Monte Carlo simulations based on a Wright–Fisher model have been performed. The results suggest that the above stochastic analysis holds only asymptotically if the product $2Ns$ is large enough (large populations or strong selection). This is an intrinsic problem of the diffusion approach, which had been noticed and discussed earlier by Stephan *et al.* (1992). With this condition satisfied, $\alpha$ depends only very weakly on the number of alleles. The corresponding values for the model with unbiased mutation rates are also given in Table II.

## 4. DISCUSSION

Genetic hitchhiking has been invoked to explain discrepancies of variability data from predictions as expected under the neutral theory (Aguadé *et al.*, 1989; Begun and Aquadro, 1991; Langley *et al.* 1993). Selected substitutions can severly reduce the variability in adjacent regions of the chromosome while being fixed in the population. The effect depends strongly on the amount of recombination between selected sites and those sites which are dragged along while the selective sweep takes place. Data of natural populations of various species have demonstrated that recombination rates and genetic variability are indeed often positively correlated (Begun and Aquadro, 1991; Begun and Aquadro, 1992; Kindahl and Aquadro, 1995; Nachmann, 1997). The hitchhiking model offers a viable explanation for this correlation: repeatedly occurring selective substitutions reduce equilibrium heterozygosity below its neutral level, and the amount of the reduction is proportional to the recombination rate. In the last few years more studies to search for correlations between the frequency of recombination in a chromosomal region and genetic variability have been performed. However, although the level of variability at microsatellites is usually far from uniform across different loci, it has, so far, been somewhat surprising to find that microsatellite variation does not display any obvious correlation with recombination rates (Michalakis and Veuille, 1996, Schloetterer *et al.*, 1997; J. Pritchard, pers. comm.). An exception is the paper by Lowenhaupt *et al.* (1989), who report some correlation between the length distribution of repeat motifs and the ability of recombination. To explain the absence of correlations it has been suggested that there is an ascertainment bias in the collection of data, where those loci which show little variability are usually disregarded in data base annotations.

The rates of mutation in motif repeat number have been determined for various species and found to be quite high; they range from about $10^{-3}$ (e.g., in *humans*, Weber and Wong, 1993) to about $10^{-6}$ (e.g., in *Drosophila melanogaster*, Schug *et al.*, 1997) mutations per locus per generation.

The above analysis shows that selective sweeps do not necessarily lead to a reduction of variability at a neighboring microsatellite locus if mutation at this locus is operating at high enough rates. However, when mutation rates are unbiased, the effect of a selective sweep is neutralized only when the mutation rate is extremely high and when the number of neutral alleles is small (see Table II). For biased mutation rates (as in (2)), the effect of a selective sweep on $V$ does not explicitly depend on the number of neutral alleles and already for moderate values of $\mu$ the hitchhiking effect is considerably weakened (see Table II and Fig. 1). Thus, selective substitutions, even when highly favorable, may just not be strong enough to override the mutation process at a linked locus during the selective phase. Figure 1 shows the admissible area (shaded in gray) of recombination and mutation rates, in order for a selective sweep to cause a reduction of 10% or more in the variance of the allele distribution (compared to the neutral value). The plot also shows that, even with complete linkage ($r = 0$), there is less than 10% reduction of $V$ if the mutation rate exceeds a maximal value $\mu^*$. More precisely, for the biased mutation model

$$\mu^* = \frac{s \log(0.1)}{8 \log(\varepsilon)} = -\frac{\log(0.1)}{4t_1}.$$

For example, $\mu^* = 2.91 \times 10^{-4}$ ($N = 10^4$) and $\mu^* = 1.98 \times 10^{-4}$ ($N = 10^6$), if $s = 0.01$. Note, that $\mu^*$ depends linearly on $s$.

The theoretical analysis uses the quantity $V$ (variance in allele size at a SSR locus) as a measure to detect the effect of selective sweeps. However, one might argue, that although selective sweeps may go undetected in terms of $V$, other quantities could be better suited to track the effect of hitchhiking events. To account for this possibility the effect on the statistic "frequency of the most frequent allele" has been checked with the help of computer simulations. One expects that a selective sweep would, on average, cause an excess in frequency of the most frequent allele over its neutral value. The results (not shown) are in perfect agreement with those based on the quantity $V$.

The presented theoretical and simulational results offer an answer, quite different from the one given before, to why the absence of correlation between the amount of recombination and the allele distribution at microsatellite loci is not surprising. The above analysis also implies that a model of hitchhiking which assumes that all neutral variation is wiped out at the end of a selective phase (Slatkin, 1995) is not adequate when mutation is present at moderate to high rates.

The results about the effect of a single hitchhiking event help to tentatively answer the question whether the equilibrium allele distribution under recurrent selective sweeps looks different from that expected under neutrality.

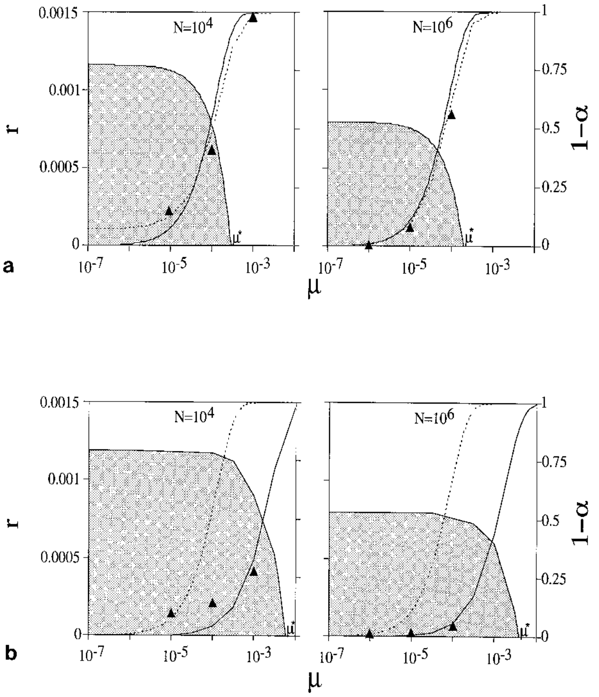

**FIG. 1.** (a) Biased mutation rates. For parameter combinations of $\mu$ (abscissa) and $r$ (recombination rate, left ordinate) within the shaded area the variance of the allele distribution at locus $\mathscr{A}$ is reduced by 10% or more compared to its neutral equilibrium value (the curve which delimits the shaded area is based on Eq. (15)). Outside of the shaded area the reduction caused by one hitchhiking event is less than 10%. Right ordinate: $1-\alpha$, which corresponds to the maximal reduction (as a fraction of the neutral value) in the zero-recombination case. Solid lines: deterministic equation (14), dotted lines: diffusion approach, according to Eq. (23). Filled triangles: simulation results, average based on 1000 replicates. The stochastic curve intersects the deterministic one and is above the latter for small mutation rates. The reason is that $V_0$ becomes small with decreasing $\mu$ in the stochastic case, but it is independent of $\mu$ in the deteministic model. To compare the simulation results to the theoretical values $\varepsilon = 1/(2N)$ had been chosen, since allele $B$ is introduced in a single copy at time $t_0$. The graphs are for $n = 10$ and $s = 0.01$. (b) Unbiased mutation rates. Abscissa and ordinates have the same meaning as in (a). However, $1-\alpha$ depends on $v$. The solid line, the shaded area and the triangles are for $n = 10$. For comparison, the dotted line shows the case of two alleles. In both cases the graphs are obtained numerically and simulation results (filled triangles) are based on 1000 replicates. Selection coefficient for all plots: $s = 0.01$.

There are two sources to weaken the effect of recurrent substitutions: mutation, in concert with recombination, opposes the reduction of $V$ below its neutral equilibrium value, while a hitchhiking event takes place. To each, the above theory applies. Second, during the neutral transitory periods; between successive selective sweeps, mutation acts to restore the neutral equilibrium, thus driving $V$ back to its value $V_0$. Therefore, crucial for the effect of recurrent hitchhiking events is how the rate of selective sweeps, $\lambda$, compares to the mutation rate. A complete analytical treatment is difficult. To derive a crude estimate, note that a lower bound to the expected time for the neutral one locus system to reach its equilibrium is $O(\mu^{-1})$ (and is independent of population size). This can be seen from the two-allele diffusion equation (see also Ewens, 1979, p. 82). Thus, the effect of recurrent selective sweeps which occur with a much lower rate than what is the neutral mutation rate will, in terms of $V$, hardly be detectable. On the other hand, if $\lambda \gg \mu$, the equilibrium allele distribution might deviate substantially from neutrality. In addition to the magnitude of the recombination, mutation and selection coefficients, also the kind of the mutation mechanism—whether rates are biased or not—plays a role. Based on compute simulations, the allele distribution under recurring hitchhiking events has been compared to the neutral reference case. Figure 2 shows results for some parameter values. The strength of the effect is positively correlated with population size (when holding the product $N\mu$ constant). This corresponds to what one would expect, since the same is true for single hitchhiking events (see Fig. 1). Furthermore, and also in accordance with expectation, the effect on $V$ is stronger under the model of unbiased mutation rates than under that of biased rates.

When accepting the view that hitchhiking occurs in natural populations, then the experimentally observed levels of $V$ and a missing correlation of microsatellite variation with recombination rates are more easily compatible with the idea that the mutation rates at SSR-sites may not be unbiased across alleles. In any case, one should be aware of the possibility that traces of hitchhiking are unlikely to be detected by statistical means, in particular when population sizes are small and mutation rates large —as in human populations ($N \approx 10^4$, $\hat{\mu} \approx 10^{-3}$).

It is left to further investigations whether the hypothesis of hitchhiking can be more distinctively rejected or established by other means than the ones discussed here —for example comparing allele distributions at homologous loci in different subpopulations of the same species. Schloetterer *et al.* (1997) recently argued in favor of this possibility and their results of a survey of *D. melanogaster* populations suggest that local and recent hitchhiking events may very well be detectable. If hitchhiking plays a role then it is conceivable that large shifts of the entire allele distribution could be accomplished in much shorter time than under mutation and drift alone. Some data of the allele distributions at SSR loci in subdivided populations of *D. melanogaster* (Michalakis and Veuille, 1996) could be interpreted in this vein. This question will be pursued elsewhere.

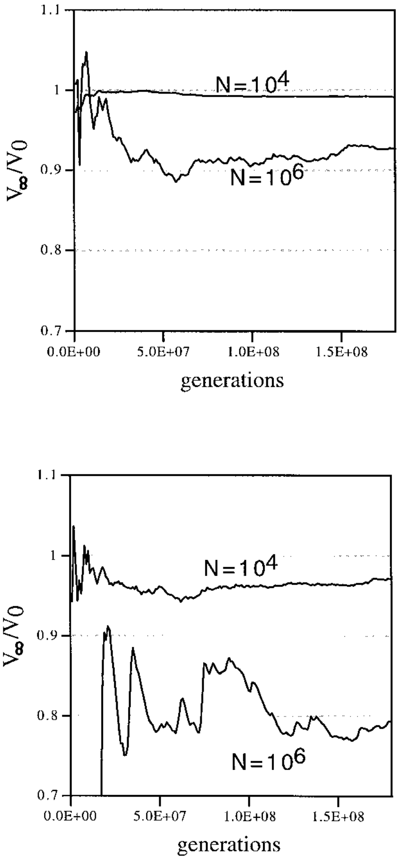

**FIG. 2.** Time average of $V_\infty/V_0$ (i.e., variance under recurrent hitchhiking events compared to variance under neutrality). The simulation technique is described in the Appendix. Parameters: $n = v + 1 = 10$, $s = 0.01$, $\lambda = 10^{-6}$, $\mu = 10^{-4}$ (case $N = 10^4$), and $\mu = 10^{-6}$ (case $N = 10^6$). Upper figure: biased mutation rates; lower figure: unbiased mutation rates.

## APPENDIX

For a biallelic neutral locus (i.e., $v = 1$), the system (3) to (5) can be solved analytically. In particular, an analytical expression for $\alpha$ can be derived. Equation (5) is decoupled and the solution given by Eq. (6). The solution of the other two equations is obtained by d'Alembert's method. They are

$$y_{1|B}(t) = \frac{\mu_0}{\mu_0+\mu_1} + \left( y_{1|B}(t_0) - \frac{\mu_0}{\mu_0+\mu_1} \right) e^{-(\mu_0+\mu_1)\,t} + (y^*_{1|B}(t) - y_{1|B}(t_0))\, e^{-(\mu_0+\mu_1)t}, \qquad \text{(A.1)}$$

$$y_{1|b}(t) = y_{1|B}(t) + (y_{1|b}(t_0) - y_{1|B}(t_0))\, e^{-(\mu_0+\mu_1+r)t}, \qquad \text{(A.2)}$$

where $\mu_0$ is the mutation rate from $A_0$ to $A_1$, $\mu_1$ the mutation rate from $A_1$ to $A_0$, and

$$y^*_{1|B}(t) = y_{1|B}(t_0) - r(y_{1|B}(t_0) - y_{1|b}(t_0)) \times \int_0^t \frac{(1-\varepsilon)\, e^{-(s+r)\tau}}{\varepsilon + (1-\varepsilon)\, e^{-s\tau}}\, d\tau$$

is the solution of $y_{1|B}$ for the case without mutation. To determine α, the ratio $V(t_1)/V(t_0)$ needs to be calculated. Since $x_B(t_0) = \varepsilon$ and $x_B(t_1) = 1-\varepsilon$, one may approximate $V(t_0)$ by $y_{1|b}(t_0)(1 - y_{1|b}(t_0))$ and $V(t_1)$ by $y_{1|B}(t_1) \times (1 - y_{1|B}(t_1))$. On doing so and taking the average over the possible initial conditions ($y_{1|B}(t_0) = 1$ is realized with probability $y_{1|b}(t_0)$ and $y_{1|B}(t_0) = 0$ is realized with probability $1 - y_{1|b}(t_0)$), one obtains

$$\frac{V(t_1)}{V(t_0)} = 1 - \varepsilon^{8\mu/s}\varepsilon^{2r/s}, \qquad \text{(A.3)}$$

where $\mu_1 = \mu_2 = \mu$ had been assumed. Furthermore, to arrive at the latter equation the approximation

$$2r \int_{t_0}^{t_1} \frac{e^{-(r+s)\tau}}{\varepsilon + e^{-s\tau}}\, d\tau \approx 1 - \varepsilon^{2r/s},$$

as justified by Stephan *et al.* (1992), had been used.

*Description of the simulations.* Simulations of recurrent selective substitutions were performed according to a Wright-Fisher model. At each time there are at most two distinct $\mathscr{B}$-alleles present in the population. At random times (according to a renewal process with parameter $\lambda$) a new selected substitution, with a fixed selective advantage $s$, is introduced into the population at frequency $x_B = \varepsilon = 1/(2N)$ ($N$ is the population size). The recombination rate between the two loci is a random variable drawn from a uniform distribution on the interval $[0; r_{\max}]$; this way selective sweeps at a distance from $\mathscr{A}$ which is larger than a maximal distance are excluded. In the course of one generation the haplotype frequencies are altered due to recombination, selection and mutation and after that the next generation is generated by multinomial sampling. A new substitution is introduced only, if the previous one had been completely fixed (generally, this restriction is not a problem as long as $1/\lambda$ is much larger than the time to fixation ($-2/s \log(\varepsilon)$), which, in the cases here, is satisfied). If a newly introduced selected substitution gets lost due to drift, then a new one is introduced immediately. The simulations are continued for about $2\ 10^8$ generations. During this period the variance, averaged over time, of the allele distribution at locus $\mathscr{A}$ is updated after each generation and recorded on an output file.

To compare this variance to the neutral case, simulations of the one-locus model (with the same parameters, except those for the linked second locus) have been carried out until the same amount of time had elapsed. When a simulation was started the deterministic equilibrium distribution (binomial and uniform, respectively) was taken as initial distribution. Average variance of the allele distribution was recorded and updated after each generation and the final variance was compared (ratio of the two variances) to the case with selective sweeps present. This result is plotted in Fig. 2 for several parameters. As estimate for $\lambda$ the one obtained by Wiehe and Stephan (1993) for *Drosophila melanogaster* has been used. Based on nucleotide variability, they estimated the index of selective sweep intensity (i.e., the product of $2Ns$ and the rate of selective sweeps) to be approximately $5.4 \times 10^{-8}$. Since this is a per site rate one has to multiply this by the maximal number of sites at which hitchhiking could be effective. The maximal recombinational distance (for $N = 10^6$) had been estimated (Wiehe and Stephan, 1993) as $r_{\max =}$ 0.002, which is $s/5$, if $s = 0.01$. Assuming that a recombinational distance of $0.2cM$ corresponds to $2 \times 10^5$ sites, the multiplier for the above rate is 20. Therefore, $\lambda = 1.08 \times 10^{-6}$.

## ACKNOWLEDGMENTS

Support of this work by the German Academic Exchange Service (DAAD) and by a grant from the U.S. NIH (GM40282) to Montgomery Slatkin is gratefully acknowledged. I would like to thank C. Garza, Y. Michalakis, R. Nielsen, and M. Slatkin for helpful comments and discussions and J. Huelsenbeck for sharing his expertise on questions of layout.

## REFERENCES

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

Akin, E. 1979. "The Geometry of Population Genetics," Lect. Notes Biomathematics, Vol. 31, Springer-Verlag, New York.

Amos, W., and Rubinstzein, D. C. 1996. Microsatellites are subject to directional evolution, *Nat. Gen.* **12**, 13–14.

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

Begun, D., and Aquadro, C. F. 1992. Levels of naturally occuring DNA polymorphism are correlated with recombination rates in *Drosophila melanogaster*, *Nature* **356**, 519–520.

Bell, G. I. 1996. Evolution of simple sequence repeats, *Computers Chem.* **20**, 41–48.

Crow, J. F., and Kimura, M. 1970. "An Introduction to Population Genetics Theory," Harper & Row, New York.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M., and Freimer, N. B. 1994. Mutational processes of simple-sequence repeat loci in human populations, *Proc. Natl. Acad. Sci. USA* **91**, 3166–3170.

Ellegren, H., Primmer, C. R., and Sheldon, B. C. 1995. Microsatellite "evolution:" Directionality or bias? *Nat. Gen.* **11**, 360–362.

Ewens, W. J. 1979. "Mathematical Population Genetics," Springer-Verlag, Berlin.

Feldman, M. W., Bergman, A., Pollock, D. D., and Goldstein, D. B. 1997. Microsatellite genetic distances with range constraints: Analytic description and problem of estimation, *Genetics* **145**, 207–216.

Freimer, N. B., and Slatkin, M. 1996. Microsatellites: evolution and mutational processes, *in* "Variation in the human genome (Ciba Foundation Symposium 197)" (D. Chadwick and G. Cardew, Eds.), pp. 51–72, Wiley, New York.

Garza, J. C., Slatkin, M., and Freimer, N. B. 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size, *Mol. Biol. Evol.* **12**, 594–603.

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

Kindahl, E. C. and Aquadro, C. F. 1995. Levels of DNA variation are correlated with rates of recombination across the third chromosome in *Drosophila melanogaster*, *Genetics* (in press).

Langley, C. H., MacDonald, J., Miyashita, N., and Aguadé, M. 1993. Lack of correlation between interspecific divergence and intra-specific polymorphism at the *suppressor of forked* region in *Drosophila melanogaster* and *Drosophila simulans*, *Proc. Natl. Acad. Sci. USA* **90**, 1800–1803.

Lowenhaupt, K., Rich, A., and Pardue, M. L. 1989. Nonrandom distribution of long mono- and dinucleotide repeats in *Drosophila* chromosomes: Correlations with dosage compensation, hetero-chromatin, and recombination, *Mol. Cell. Biol.* **9**, 1173–1182.

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

Michalakis, Y., and Veuille, M. 1996. Length variation of CAG/CAA trinucleotide repeats in natural populations of *Drosophila melanogaster* and its relation to the recombination rate, *Genetics* **143**, 1713–1725.

Nachman, M. W. 1997. Patterns of DNA variability at *X*-linked loci in *Mus domesticus*, *Genetics* **147**, 1303–1316.

Ohta, T., and Kimura, M. 1975. The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect), *Genet. Res.*, *Camb.* **25**, 313–326.

Press, W. H., Teukolsky, S. A., Brian, P., and Flannery, W. T. V. 1992. "Numerical Recipes in C," Cambridge University Press, Cambridge.

Schloetterer, C., Vogl, C., and Tautz, D. 1997. Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations, *Genetics* **146**, 309–320.

Schug, M. D., Mackay, T. F. C., and Aquadro, C. F. 1997. Low mutation rates of microsatellite loci in *Drosophila melanogaster*, *Nature Genetics* **15**, 99–102.

Slatkin, M. 1995. Hitchhiking and associative overdominance at a microsatellite locus, *Mol. Biol. Evol.* **12**, 473–480.

Stephan, W., Wiehe, T., and Lenz, M. W. 1992. The effect of strongly selected substitutions on neutral polymorphism: Analytical result based on diffusion theory, *Theor. Pop. Biol.* **41**, 237–254.

Valdes, A. M., Slatkin, M., and Freimer, N. B. 1993. Allele frequencies at microsatellite loci: The stepwise mutation model revisited, *Genetics* **133**, 737–749.

Weber, J. L., and Wong, C. 1993. Mutation of human short tandem repeats, *Hum. Mol. Gen.* **2**, 1123–1128.

Wiehe, T., and Stephan, W. 1993. Analysis of a genetic hitchhiking model and its application to DNA polymorphism data from *Drosophila melanogaster*, *Mol. Biol. Evol.* **10**, 842–854.

Zhivotovsky, L. A., and Feldman, M. W. 1995. Microsatellite variability and genetic distances, *Proc. Natl. Acad. Sci. USA* **92**, 11549–11552.