# Constrained Disequilibrium Values and Hitchhiking in a Three-Locus System

**Mark N. Grote,* William Klitz† and Glenys Thomson‡**

**Section of Evolution and Ecology, University of California, Davis, California 95616, †School of Public Health, University of California, Berkeley, California 94720 and ‡Department of Integrative Biology, University of California, Berkeley, California 94720*

Manuscript received February 6, 1998
Accepted for publication August 7, 1998

ABSTRACT

Positive selection on a new mutant allele can increase the frequencies of closely linked alleles (through hitchhiking), as well as create linkage disequilibrium between them. Because this disequilibrium is induced by the selected allele, one may be able to identify loci under selection by measuring the influence of a candidate locus on pairwise disequilibrium values at nearby loci. The constrained disequilibrium values (CDV) method approaches this problem by examining differences in pairwise disequilibrium values, which have been normalized for two- and three-locus systems, respectively. We have investigated in detail the reliability of inferences based on CDV, using simulation and analytical methods. Our main results are (i) in some circumstances, CDV may not distinguish well between a selected locus and a neighboring neutral locus, but (ii) CDV seldom indicates "selection" in neutral haplotypes with moderate to large $4Nc$. We conclude that, although the CDV method does not appear to precisely locate selected alleles, it can be used to screen for regions in which hitchhiking is a plausible hypothesis. We present a microsatellite data set from human chromosome 6, in which constrained disequilibrium values suggest the action of selection in a region containing the human leukocyte antigen (HLA)-A and myelin oligodendrocyte glycoprotein (MOG) loci. The connection between hitchhiking and disequilibrium has received relatively little attention, so our investigation presents opportunities to address more general issues.

IN the genetic hitchhiking model, positive selection on a new mutant allele increases the frequencies of other alleles physically linked to the mutant, skewing the frequency distributions at the linked loci. Theoretical and empirical studies of hitchhiking generally focus on the reduction in variation at linked neutral loci that can result if the recombination rate is low and the selected mutant is quickly fixed in the population (Maynard-Smith and Haigh 1974; Ohta and Kimura 1975; Aguadé *et al.* 1989; Kaplan *et al.* 1989; Stephan and Langley 1989; and many more recent studies). Relatively fewer studies have focused on linkage disequilibrium (gametic phase disequilibrium) in haplotypes subject to hitchhiking (Thomson 1977; Robinson *et al.* 1991a; Begun and Aquadro 1994, 1995). Our concerns in the present study are the nature of linkage disequilibrium created by hitchhiking, and the extent to which certain patterns of disequilibrium can be used to make inferences about hitchhiking. In a wider sense, our aim is to investigate a particular method that uses linkage disequilibrium to physically locate genes of interest.

Relatively insignificant linkage disequilibrium is always created by the appearance of a new mutant, because initially the mutant is found only on an "ancestral" haplotype of closely linked alleles. Thomson (1977) showed that hitchhiking can noticeably increase this existing disequilibrium, if selection in favor of the new mutant is strong enough relative to the recombination rate with linked loci. Hitchhiking can also create significant disequilibrium between nonselected alleles if they are closely linked to the selected allele (Thomson 1977); here, disequilibrium between neutral alleles is induced by mutual association with the selected allele. These associations are expected to decline in strength as recombination breaks up haplotypes bearing the selected mutant.

Robinson *et al.* (1991a,b) introduced the constrained disequilibrium values (CDV) method as a means of identifying loci that may have been subject to recent hitchhiking. Inference with the CDV method depends on comparisons between pairwise disequilibrium measures, which have been normalized according to different constraints imposed by two- and three-locus systems. A familiar two-locus linkage disequilibrium measure is

$$D_{ab} = f_{ab} - p_a p_b = f_{ab} f_{AB} - f_{aB} f_{Ab}, \qquad (1)$$

where $f_{ab}$ is the frequency of the $ab$ haplotype and $p_a$, $p_b$ are the corresponding one-locus allele frequencies. The value of $D_{ab}$ depends strongly on the magnitudes of $p_a$ and $p_b$, so $D_{ab}$ is commonly normalized using upper and lower bounds imposed by $p_a$ and $p_b$ (Lewontin 1964, 1988; Hedrick 1987). Robinson *et al.* (1991b) showed that a third locus imposes further bounds on $D_{ab}$, and showed how to normalize $D_{ab}$ using these additional constraints (described below). Differences in pairwise disequilibria, normalized in the two different ways, high-

*Corresponding author:* Mark N. Grote, Section of Evolution and Ecology, University of California, Davis, CA 95616.
E-mail: mngrote@ucdavis.edu

light the influence that a third locus may exert on the pairwise measure. On the basis of deterministic simulations of a three-locus hitchhiking model, Robinson *et al.* (1991a) proposed that differences in the normalized measures could indicate which of the three loci has the selected mutant. The method for making such inferences was termed the CDV method.

Our purpose is to present some recent results that bear upon the use and interpretation of CDV. First, we summarize further simulations of the deterministic model, describing some circumstances under which CDV does, or does not, lead to reliable inferences about the position of the selected locus. In connection with this, we analyze the normalized disequilibrium measures under a selection model with simplifying assumptions and show that inferences with CDV are especially sensitive to allele frequencies at neutral loci closely linked to the selected locus. Second, we apply the CDV method to data sets generated under a stochastic model of neutral haplotypes, using a simulation program of Hudson (1983, 1985), to assess the performance of CDV under a finite-population "null" model. In our discussion, we reexamine the types of inferences that can be made with CDV and address some conceptual and practical issues. Finally, we apply the CDV method to marker haplotypes from human chromosome *6*, to illustrate one use of CDV.

## METHODS

**Measures of disequilibrium and hitchhiking:** Our attention centers on two normalized measures of pairwise linkage disequilibrium, $D'$ and $D''$, and in particular on the difference in their magnitudes,

$$\delta = |D'| - |D''|.$$

$D'$ is the familiar normalized pairwise linkage disequilibrium measure (Lewontin 1964, 1988; Hedrick 1987; Robinson *et al.* 1991b), where for alleles $a$ and $b$ at distinct loci,

$$D'_{ab} = \begin{cases} \dfrac{D_{ab}}{\min(p_a q_b,\ q_a p_b)} & \text{if } D_{ab} > 0 \\ 0 & \text{if } D_{ab} = 0 \\ \dfrac{D_{ab}}{-\max(-p_a p_b,\ -q_a q_b)} & \text{if } D_{ab} < 0, \end{cases} \tag{2}$$

with $D_{ab}$, $p_a$, and $p_b$ as in (1), $q_a = 1 - p_a$ and $q_b = 1 - p_b$. $D'_{ab}$ is calculated from $D_{ab}$ via division by an appropriate upper or lower bound. The extreme values of $D'_{ab}$ reflect the complete association of the $a$ or $b$ allele with the $ab$ haplotype ($D'_{ab} = +1$), or the absence of any $ab$ haplotypes when in fact the constituent alleles are present ($D'_{ab} = -1$).

Robinson *et al.* (1991b) derived a new normalized pairwise measure, $D''_{ab}$, which incorporates futher constraints imposed on $D_{ab}$ by a third locus. For a three-locus diallelic haplotype, where the C locus plays the role of the "constraining" locus,

$$D''_{ab(c)} = \begin{cases} \dfrac{D_{ab}}{\max^* D_{ab}} & \text{if } D_{ab} > 0 \text{ and } \min^* D_{ab} \leq 0 \\ \dfrac{D_{ab} - \min^* D_{ab}}{\max^* D_{ab} - \min^* D_{ab}} & \text{if } D_{ab} > 0 \text{ and } \min^* D_{ab} > 0 \\ 0 & \text{if } D_{ab} = 0 \\ \dfrac{D_{ab}}{-\min^* D_{ab}} & \text{if } D_{ab} < 0 \text{ and } \max^* D_{ab} \geq 0 \\ \dfrac{D_{ab} - \max^* D_{ab}}{\max^* D_{ab} - \min^* D_{ab}} & \text{if } D_{ab} < 0 \text{ and } \max^* D_{ab} < 0, \end{cases} \tag{3}$$

where

$$\min^* D_{ab} = \max(-p_a p_b,\ -q_a q_b,\ -m_1,\ -m_2)$$
$$\max^* D_{ab} = \min(p_a q_b,\ q_a p_b,\ M_1\ M_2)$$

and

$$m_1 = p_a p_b p_c + q_a q_b q_c + D_{ac} + D_{bc}$$
$$m_2 = p_a p_b q_c + q_a q_b p_c - D_{ac} - D_{bc}$$
$$M_1 = p_a q_b p_c + q_a p_b q_c + D_{ac} - D_{bc}$$
$$M_2 = p_a q_b q_c + q_a p_b p_c - D_{ac} + D_{bc}.$$

The associations between $a$ and $c$, and between $b$ and $c$, enter the calculation of $D''_{ab}$ through $m_1$, $m_2$, $M_1$, and $M_2$.

Like $D'_{ab}$, $D''_{ab}$ lies between $+1$ and $-1$, where the extreme values indicate the strongest possible positive or negative association between alleles $a$ and $b$, within the constraints imposed by the allele frequencies and pairwise disequilibria of the three-locus system. We write $D''_{ab(c)}$ because $D''_{ab}$ is calculated with reference to a particular allele at the third locus, but in a diallelic system, one can show $D''_{ab(c)} = D''_{ab(C)}$. Moreover, as with $D'_{ab}$ in a diallelic system, $D''_{ab} = D''_{AB} = -D''_{aB} = -D''_{Ab}$.

Assuming $D_{ab} > 0$ for the moment, $\delta = |D'| - |D''|$ is greater than zero when the pairwise measure $D_{ab}$ is more extreme relative to its *two*-locus maximum, than to its positive range in the *three*-locus system; in this case the pairwise association between $a$ and $b$ appears to be relatively weaker when all of the pairwise associations of the three-locus system are taken into account. Loosely speaking, when $\delta > 0$, the association between $a$ and $b$ is said to be partly accounted for by their mutual association with $c$. Assuming further that $c$ is a selected mutant, this property of $\delta$ is the primary reason for treating $\delta > 0$ as the "footprint" of a hitchhiking event, in which the neutral $a$ and $b$ alleles have hitchhiked with $c$ (Robinson *et al.* 1991a).

Although the normalized measure $D'$ may change during a hitchhiking event (Thomson 1977), $D'$ alone does not distinguish between loci that may be under positive selection and linked neutral loci that are only hitchhiking with the selected locus. Similarly, there is a

single measure of third-order disequilibrium, $D_{abc}$, which can also be normalized appropriately (Geiringer 1944; Thomson and Baur 1984), but $D_{abc}$ also makes no distinction between selected and hitchhiking loci. The main claim of Robinson *et al.* (1991a) is that δ values, when interpreted appropriately, can make this distinction.

For a given three-locus haplotype, each locus may play the role of the constraining locus, and there are three δ values: $\delta_{ab(c)}$, $\delta_{a(b)c}$, and $\delta_{(a)bc}$. Using deterministic simulations, Robinson *et al.* (1991a) found that δ was often large and positive when the "constraining" allele was increasing in frequency due to positive selection, but the linked alleles were selectively neutral. When a nonselected allele played the "constraining" role, Robinson *et al.* (1991a) found that δ tended to be zero or negative. Based on their observations, Robinson *et al.* (1991a) proposed the following criteria for inferring selection based on δ values:

1. If one of the three δ values is positive and the remaining two are zero or negative, the constraining allele that gives the positive δ is the one that may have experienced recent selection.
2. If more than one of the δ values is positive, but one is much larger than the rest (for this study, more than double the next largest), the constraining allele that gives the large δ is the one that may have experienced recent selection.
3. If all three δ values are ≤0, or two are positive but close in value, no conclusion about selection can be drawn.

Robinson *et al.* (1991a) paid considerable attention to the magnitudes of δ values under various scenarios, but we first focus simply on which loci the CDV method identifies as candidates for selection, in a large series of deterministic simulations.

**A deterministic hitchhiking model:** The deterministic simulations are based on a three-locus, diallelic model that evolves via a standard system of algebraic recursions (Feldman *et al.* 1974; Thomson 1977; Hartl and Clark 1989). For purposes of the CDV method, we are interested in a single new mutant allele and the closely linked alleles of the ancestral haplotype on which the mutant first appeared. The alleles of interest, in their order on the chromosome, are *a*, *b*, and *c*, one of which will be the new mutant and the others linked alleles. By convention, *A*, *B*, and *C* may be taken to represent all other alleles at their respective loci.

The recursion equations describing changes in the haplotype frequencies can be specified by selection and mutation parameters described immediately below, the recombination rates $r_1$ and $r_2$ between the A and B loci and the B and C loci, respectively (where $r_1 + r_2 - 2r_1r_2$ gives the recombination rate between A and C for the "no-interference" model), and a set of initial haplotype frequencies. The latter are determined by specifying initial allele frequencies $p_a(0)$, $p_b(0)$, $p_c(0)$, and a single initial disequilibrium value [*e.g.*, $D'_{ab}(0)$ when *c* is the new mutant]. In addition, we assume that the haplotype bearing the new mutant has not experienced mutation or recombination before the simulation begins at generation zero [for example, if *c* is the new mutant, this implies $f_{abc}(0) = p_c(0)$]. The frequency dynamics of a strongly selected allele, once it has left the zero-frequency boundary, are commonly modeled as a deterministic process (*e.g.*, Kaplan *et al.* 1989). For convenience, we have assumed that the time spent close to the boundary is small relative to recombination and mutation rates near the selected locus.

Fitnesses at the selected locus (using genotypes at the C locus for illustration) are given by $w_{cc} = 1 - s_c$, $w_{Cc} = 1$, and $w_{CC} = 1 - s_C$. We have adopted a general framework for hitchhiking studies, as our selection model encompasses both directional selection leading to fixation of the new mutant (*e.g.*, $s_c \leq 0$ and $0 < s_C \leq 1$) and balancing selection ($0 < \{s_C, s_c\} < 1$). Mutation is unidirectional at rates $\mu_a = \mu_b = \mu_c = 10^{-5}$ per generation from alleles *a*, *b*, and *c* to *A*, *B*, and *C*, respectively, so that the alleles of interest are transient. We use terms like "equilibrium frequency" loosely, referring to the relatively fast adjustment of allele frequencies that results from the appearance of a new selected mutant. For completeness, we have included the recursion equations in the appendix.

**Scope of the deterministic simulations:** The parameter space for the deterministic model is large and multidimensional, so we limit our investigation to a relatively narrow subset of parameter values under which measurable linkage disequilibrium is likely to be present. Using simple frequency arguments, one can conclude that most new mutants arise on relatively common haplotypes; but more unusual events, in which mutants appear on rare haplotypes, are actually of greater interest in hitchhiking studies. Thomson (1977) showed that hitchhiking will only noticeably perturb allele frequencies and disequilibria when at least one of the neutral alleles initially linked to the mutant is rare. The pairwise disequilibrium value $D_{ab}$ is only large when alleles *a* and *b* are at intermediate frequencies and strongly associated in the "coupling" (*ab*) phase. An initially rare *ab* haplotype, on which a strongly selected mutant happens to occur, is in a primary position to pass through this range of intermediate frequencies in strong coupling.

In the following simulations, we have (somewhat arbitrarily) set the initial frequency of at least one of the neutral alleles at $p(0) = 0.05$, to ensure that the ancestral haplotype is sufficiently rare. Table 1 shows parameter values that are typical of the simulations. Here, *c* is the selected mutant and $p_a(0)$ and $p_b(0)$ are treated in a symmetric fashion, each assuming the value $p(0) = 0.05$ while the other takes values between 0.05 and 0.9 in successive runs. Some values of the initial pairwise disequilibrium $D'_{ab}(0)$ rule out certain combinations of

$p_a(0)$, $p_b(0)$, and $p_c(0)$ in Table 1, but the same treatments are always applied to the *a* and *b* alleles.

Values of the remaining parameters were guided by a few basic rules. Hitchhiking is thought to be a weak force unless selective values are roughly an order of magnitude greater than recombination rates (Maynard-Smith and Haigh 1974; Thomson 1977; Kaplan *et al.* 1989), so we have chosen selection and recombination parameters accordingly. Generally, for each setting of $p_a(0)$, $p_b(0)$, $p_c(0)$, $D'_{ab}(0)$, $r_1$, and $r_2$ that was investigated, we examined a basic series of runs formed by $7 \times 10 = 70$ pairs of selection coefficients (for example, $s_C$ and $s_c$ as shown in Table 1). Combinations of parameter values that involved interactions beyond those of primary interest [for example, $D'_{ab}(0) = -0.25$ and $r_1 \neq r_2$ in Table 1] were left unexamined to keep the number of runs reasonable. More detailed tables are in Grote (1996) and are available upon request.

Within these guidelines, our first objectives are to significantly enlarge upon the number of deterministic cases examined in Robinson *et al.* (1991a), and to investigate some cases where the relationship between the δ values is inconsistent with correct inference of the selected locus.

**CDV in a stochastic neutral model:** Our second aim is to study the performance of CDV in a neutral, finite-population model, to determine whether or not genetic drift and sampling effects can produce patterns of linkage disequilibrium conforming to criteria 1 or 2 above. Robinson *et al.* (1991a) used an ad hoc method to study δ values under genetic drift.

We have modified a computer program of Hudson (1983, 1985) to study δ values in the neutral model. The program simulates random samples of three-locus haplotypes, generated under the neutral "infinite alleles" model with recombination at equilibrium. The program requires the following input parameters: *n*, the number of haplotypes per sample; 4*Nc*, the scaled recombination rate between the A and C loci (the B locus is assumed to be halfway between A and C); $\theta_a$, $\theta_b$, and $\theta_c$ (with, *e.g.*, $\theta_a = 4N\mu_a$, where *N* is the effective population size and $\mu_a$ is the mutation rate to new A-locus alleles). We used the value $\theta = 0.2$ at each locus, corresponding to the approximate numerical solution of

$$E[K_n] = \frac{\theta}{\theta} + \frac{\theta}{(\theta + 1)} + \cdots + \frac{\theta}{(\theta + n - 1)} = 2$$

(Ewens 1979), with $E[K_n]$ the expected number of alleles (per locus) in the sample and $n = 100$. As one would expect, not all samples generated at $\theta = 0.2$ were segregating exactly two alleles at each locus, so we screened each sample and retained only those with diallelic loci. We further required in each sample a standard minimum level of heterozygosity, $H \geq 0.095$ per locus. We then calculated the three values, $\delta_{(a)bc}$, $\delta_{a(b)c}$, and $\delta_{ab(c)}$ in each accepted sample. For each of three levels of recombination 4*Nc*, we generated independent samples until 1000 samples had met the screening criteria; our stochastic simulation results are based on these groups of 1000 samples.

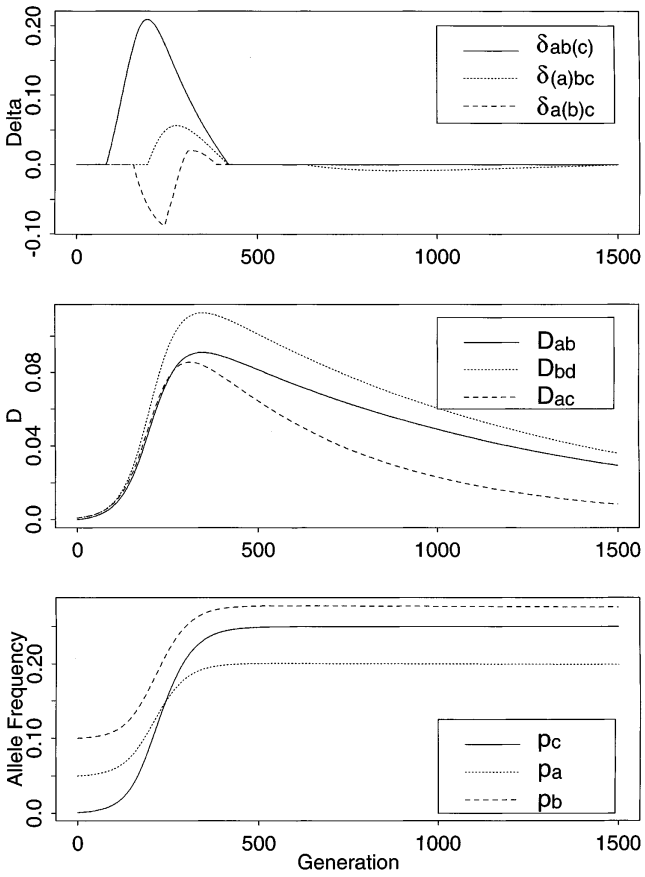


Figure 1.—Deterministic simulation for selection at *c*, with $D'_{ab}(0) = 0.0$, $r_1 = r_2 = 0.001$, $p_a(0) = 0.05$, $p_b(0) = 0.1$, $p_c(0) = 0.001$, $s_C = 0.025$, $s_c = 0.075$.

## RESULTS

**Deterministic simulations:** Figures 1–3 show sample runs of a deterministic model in which *c* is the selected mutant and the A and B loci are neutral. In Figures 1–3, recombination rates, initial allele frequencies at the A and C loci, and mutation and selection parameters are the same; only the initial frequency of the *b* allele varies between the figures.

In the allele frequency plots, $p_c(t)$ approaches the equilibrium value $s_C/(s_c + s_C) = 0.25$, then slowly declines due to mutation (not evident in these plots). Frequencies of the *a* and *b* alleles both increase due to hitchhiking with the selected mutant *c*. The frequencies ultimately attained by *a* and *b* depend on their initial frequencies, $D_{ab}(0)$, the strength of selection on *c*, and the recombination rates between these loci (Maynard-Smith and Haigh 1974; Thomson 1977). The initial

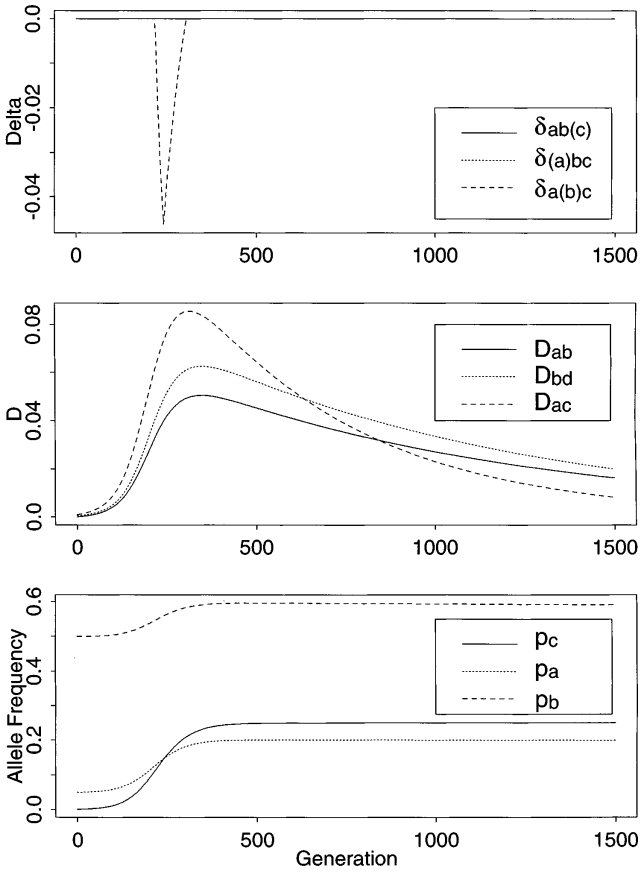


Figure 2.—Deterministic simulation for selection at *c*, the same as Figure 1 except $p_b(0) = 0.5$.

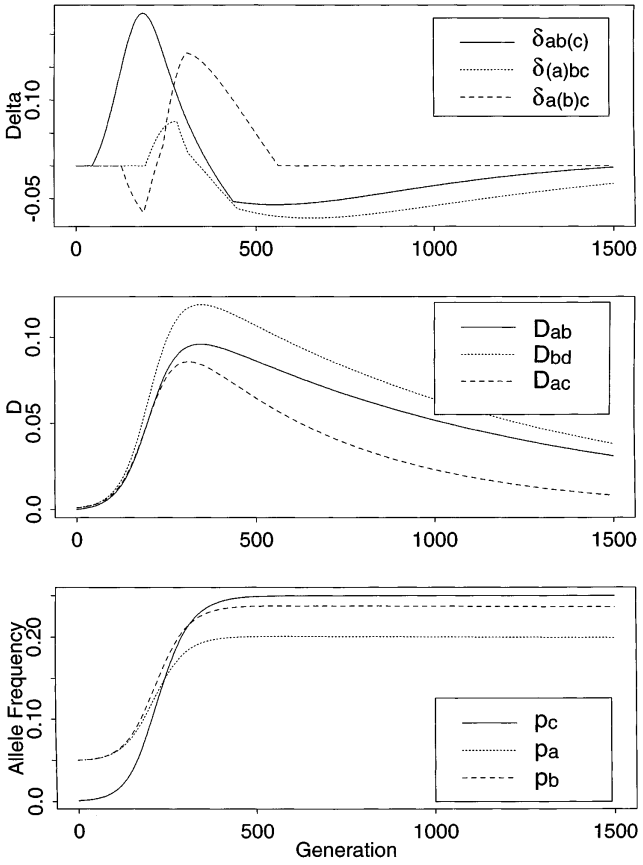


Figure 3.—Deterministic simulation for selection at *c*, the same as Figure 1 except $p_b(0) = 0.05$.

value of $D_{ab}$ is zero in each of the runs, and initial values of $D_{ac}$ and $D_{bc}$ are small positive numbers reflecting the associations of *a* and *b* with the new mutant *c*. All three values of *D* increase with the hitchhiking effect, then slowly decrease. Because there is only a single selected locus in these runs, the equilibrium value of all disequilibria *D* and each δ is zero. In the deterministic model without selection on *c*, $D_{ab}$ would remain at zero, whereas $D_{ac}$ and $D_{bc}$ would decline from their initial values to zero without a transient increase. In these runs, it is only after the disequilibrium measures *D* have attained relatively large values that deviations from $\delta = 0$ are observed.

In Figure 1, δ values between roughly generations 100 and 300 satisfy criteria 1 or 2 to correctly indicate selection at the C locus. Later in the run δ values conform to criterion 3, where no conclusions about selection would be made. Figure 2 conforms entirely to criterion 3, having no signal for selection during the run. In Figure 3, both the *b* and *c* alleles meet criteria for selection at different times in the run, although only *c* is under selection. In particular, applying the CDV criteria at any time between generations 320 and 550, we could conclude that the neutral *b* allele is in fact under positive selection (Figure 3 is similar to pattern II′ in Figure 3 of Robinson *et al.* 1991a). Inferences about the selected allele based on disequilibrium values at a single time-point could indeed be misleading in Figure 3, where knowledge of the whole history of the selective event might seem necessary for correct inference.

The performance of the CDV criteria in a large series of deterministic runs is summarized in Tables 1 and 2. Following the discussion above, we have classified each run by determining which alleles, if any, the CDV criteria would indicate as "selected." The run of Figure 1 shows a correct signal at the C locus for $100 \leq t \leq 300$ but gives no signal for selection otherwise, and is counted under the column "signal at *c* alone" in Table 1. Sampling such a run at an arbitrary time, we might draw no conclusions, but would not incorrectly identify a neutral allele as selected. The run of Figure 2 gives no signal for selection at all and is counted under "no signal" in Table 1. The run of Figure 3 gives, for $320 \leq t \leq 550$, a misleading signal for selection at the neutral B locus and is counted under "signal at *b*" in Table 1 (there is a similar column for "signal at *a*"). Because there are no runs in this series with signals at both neutral loci, each run falls into only one of these categories. We have chosen a conservative classification that emphasizes times during which CDV leads to incorrect inferences. In the text below, we describe broad

**TABLE 1**

***c* is the new mutant under selection**

| $p_c(0)$ | $D'_{ab}(0)$ | $r_1$, $r_2$ | $p(0)$: Neutral alleles | | $s_C$ | $s_c$ | Signal at *c* alone | Signal at *a* | Signal at *b* | No signal | Total runs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.001 | −0.25 | 0.001 | 0.05 | 0.05 | 0.01 | −0.05 | 676 (26.1%) | 113 (4.4%) | 1149 (44.4%) | 652 (25.2%) | 2590 |
| 0.01 | 0.0 | 0.005 | | 0.1 | 0.025 | −0.025 | | | | | |
| | 0.25 | 0.01 | | 0.3 | 0.05 | −0.01 | | | | | |
| | | | | 0.4 | 0.075 | 0.0 | | | | | |
| | | | | 0.5 | 0.09 | 0.01 | | | | | |
| | | | | 0.75 | 0.1 | 0.025 | | | | | |
| | | | | 0.9 | 0.15 | 0.05 | | | | | |
| | | | | | | 0.075 | | | | | |
| | | | | | | 0.1 | | | | | |
| | | | | | | 0.15 | | | | | |

Parameter values and performance of the CDV method in deterministic runs with selection at allele *c*. See methods for a description of the table's structure and results for details of the runs.

**TABLE 2**

***b* is the new mutant under selection: $p_c(0) = 0.05$, $r_2 = 0.001$**

| $p_b(0)$ | $D'_{ac}(0)$ | $r_1$ | $p_a(0)$ | $s_B$, $s_b$ | Signal at *b* alone | Signal at *a* | Signal at *c* | No signal | Total runs |
|---|---|---|---|---|---|---|---|---|---|
| 0.001 | −0.25 | 0.001 | 0.05 | (Same as $s_C$, $s_c$ in Table 1) | 979 (70.0%) | 0 (0%) | 194 (13.9%) | 227 (16.2%) | 1400 |
| 0.01 | 0.0 | 0.005 | 0.1 | | | | | | |
| | 0.25 | 0.01 | 0.3 | | | | | | |
| | | | 0.4 | | | | | | |
| | | | 0.5 | | | | | | |
| | | | 0.75 | | | | | | |
| | | | 0.9 | | | | | | |

Deterministic runs for selection at *b*, with the A and C loci neutral.

trends and give some breakdowns of the runs that would not be evident by examining Tables 1 and 2 alone. We use percentages in the tables and text as convenient summaries, but do not view these as probabilities.

In 26.1% (676/2590) of the runs in Table 1, the only signal identifying an allele under selection correctly points to *c* as the selected mutant. The CDV criteria identify the selected locus most reliably when the *b* allele is initially of moderate frequency: 50.5% (283/560) of the runs in Table 1 with $p_b(0) = 0.3$, 0.4, or 0.5 and $p_a(0) = 0.05$ resulted in the *c* allele being correctly identified, 35.5% (199/560) led to a possible misidentification of the *b* allele, and the remaining 14.0% (78/560) gave no signal for selection. CDV also performs well when the initial disequilibrium between the neutral alleles is negative: 46.7% (294/630) of the runs with $D'_{ab}(0) = -0.25$ correctly identified the *c* allele and only 20.8% (131/630) gave false signals at *a* or *b*. CDV performs poorly when the initial frequencies of the neutral loci differ widely, resulting in a false signal for selection at the rarer of the two neutral alleles: 52.1% (219/420) of the runs with $p_a(0) = 0.75$ or 0.9 and $p_b(0) = 0.05$ gave a false signal at *b*, and 33.3% (140/420) of the runs with $p_a(0) = 0.05$ and $p_b(0) = 0.75$ or 0.9 gave a false signal at *a*. In general, CDV does a poor job identifying the selected allele when the *b* allele is rare: 62.1% (956/1540) of the runs with $p_b(0) = 0.05$ gave a false signal for selection at the *b* allele. When the *b* allele is initially rare, the CDV criteria do not distinguish well between the new selected mutant *c* and its closest neutral neighbor.

In these simulations, when $s_c \leq 0.0$ and $s_C > 0.0$, the new mutant *c* will be transiently fixed in the population (often called a "selective sweep"). In the selective-sweep runs, 24.8% (257/1036) gave a correct signal from the *c* allele, 50.5% (523/1036) led to a possible misidentification of the *b* allele, and 4.2% (43/1036) to a possible misidentification of the *a* allele. In the next section, we will examine why CDV may not perform especially well in a selective sweep.

As one might imagine, there is a trend in the reliability of inferences associated with the ratio $s_c/s_C$: for fixed values of the remaining parameters, with both $s_c$, $s_C > 0$, runs with larger values of $s_c/s_C$ tend to have no signal, those with smaller values of $s_c/s_C$ tend to have incorrect signals from the *b* allele, and those with intermediate values of $s_c/s_C$ allow the CDV method to perform best. The critical values of the ratio $s_c/s_C$ depend in a complex way on the remaining parameters and appear to be different in each series of runs.

Table 2 is similar in structure to Table 1, except now *b* is the new selected mutant and the A and C loci are neutral. Here, by symmetry, there is no need to switch the roles of the neutral loci, and we use $p_c(0) = 0.05$, $r_2 = 0.001$ throughout. When *b* is the new mutant, a relatively small number of runs have a potentially misleading signal at a neutral locus, and nearly all are cases in which the neutral allele frequencies $p_a(0)$ and $p_c(0)$ differ widely [*i.e.*, $p_a(0) = 0.75$ or 0.9 and $p_c(0) = 0.05$].

**The role of allele frequencies at a closely linked neutral locus:** The most problematic observation in the simulations above was a strong tendency for the CDV method to indicate selection at the B locus when *c* was the new selected mutant. Using some mathematics and general aspects of the hitchhiking model, it is possible to show that a rare neutral allele on the ancestral haplotype can easily be mistaken for the selected allele, when using the CDV method. The analysis requires some simplifying assumptions, but gives some generality to the results of the deterministic simulations, showing that our observations do not depend strongly on particular choices of parameter values.

*An overdominance model:* We examine the behavior of $\delta_{a(b)c}$, the δ value that indicates selection at the B locus, during the rapid increase of a new, strongly overdominant *c* allele. To avoid dealing with the time component explicitly, we focus on $\delta_{a(b)c}$ at $t = 0$, $t$ "small" (a few generations) and $t$ "moderate" (on the order of 100 to a few hundred generations). We assume that $r_1$ and $r_2$ are small enough so that recombination in the ancestral haplotype *abc* can be practically ignored when $t$ is near zero, and further assume that $p_b(0)$ is small enough so that *b* and *c* are in strong coupling for small-to-moderate $t$. Low recombination and strong coupling of *b* and *c* imply that $f_{ab}(t)$, $f_{ac}(t)$, and $f_{bc}(t)$ are all approximately equal to $p_c(t)$ for small-to-moderate $t$. We finally assume $D_{ab} = 0$, but due to hitchhiking, all of $D_{ab}$, $D_{ac}$, and $D_{bc}$ are positive after a few generations of selection.

To characterize $\delta_{a(b)c}$, we must study the relationship between $D'_{ac}$ and $D''_{a(b)c}$ for $t = 0$, $t$ small and $t$ moderate. For convenience, the required definitions when $D_{ac} \geq 0$ are

$$D'_{ac} = \frac{D_{ac}}{\min(p_a q_c,\ q_a p_c)}$$

and

$$D''_{a(b)c} = \begin{cases} \dfrac{D_{ac}}{\max^* D_{ac}} & \text{if } \min^* D_{ac} \leq 0 \\ \dfrac{D_{ac} - \min^* D_{ac}}{\max^* D_{ac} - \min^* D_{ac}} & \text{if } \min^* D_{ac} > 0, \end{cases}$$

where

$$\min^* D_{ac} = \max(-p_a p_c,\ -q_a q_c,\ -m_1,\ -m_2)$$
$$\max^* D_{ac} = \min(p_a q_c,\ q_a p_c,\ M_1,\ M_2)$$

and

$$m_1 = p_a p_b p_c + q_a q_b q_c + D_{ab} + D_{bc}$$
$$m_2 = p_a q_b p_c + q_a p_b q_c - D_{ab} - D_{bc}$$
$$M_1 = p_a p_b q_c + q_a q_b p_c + D_{ab} - D_{bc}$$
$$M_2 = p_a q_b q_c + q_a p_b p_c - D_{ab} + D_{bc}.$$

Because all of $D_{ab}$, $D_{ac}$, and $D_{bc}$ are $\geq 0$ by assumption,

the sign of min*$D_{ac}$ is determined entirely by the relative sizes of the positive and negative terms in $m_2$. When the disequilibria $D_{ab}$ and $D_{bc}$ in $m_2$ are small relative to the third-order products of allele frequencies (as they will tend to be for $t$ near zero), $m_2 \geq 0$ and min*$D_{ac} \leq 0$. When the disequilibria are large relative to the third-order products (as they tend to be for moderate $t$), $m_2 < 0$ and min*$D_{ac} > 0$.

At $t = 0$, $f_{ac} = p_c$, and because the new mutant $c$ is found only with $a$, $D_{ac} = \max D_{ac}$. Further, the inequality

$$D_{ac} \leq \max^* D_{ac} \leq \max D_{ac}$$

must hold, because the set ($p_a q_c$, $q_a p_c$, $M_1$, $M_2$) that determines max*$D_{ac}$ contains the set that determines max $D_{ac}$. $D_{ac} = \max D_{ac}$ then implies max*$D_{ac} = \max D_{ac}$, and therefore

$$\delta_{a(b)c} = \frac{D_{ac}}{\max D_{ac}} - \frac{D_{ac}}{\max^* D_{ac}} = 0$$

for $t = 0$.

For small $t$, with the disequilibria of $m_2$ still small relative to the third-order products, the reasoning is very similar. Because the new mutant $c$ is still found almost exclusively on the ancestral haplotype, $D_{ac} \approx \max D_{ac}$ to a good approximation, so it also must be true that max*$D_{ac} \approx \max D_{ac}$. We then have $\delta_{a(b)c} \approx 0$ for small $t$.

The situation changes when the disequilibria of $m_2$ are large relative to the third-order products, so that $m_2 < 0$ and min*$D_{ac} > 0$; here, we must use the second case in the definition of $D''_{ac}$ above. We further observe that when the loci are evenly spaced, recombination begins relatively soon to reduce $D_{ac}$ below its two-locus maximum (compared to $D_{ab}$ and $D_{bc}$), although all of the disequilibria may have dropped below earlier large values due to allele frequency constraints. Now consider $t$ moderate, with min*$D_{ac} > 0$ and $D_{ac} < \max D_{ac}$. To determine the sign of $\delta_{a(b)c}$, we must examine as before the relative magnitudes of max $D_{ac}$ and max*$D_{ac}$. It is convenient to use algebraically equivalent expressions for the terms $M_1$ and $M_2$ in max*$D_{ac}$:

$$M_1 = q_a p_c + f_{ab} - f_{bc}$$
$$M_2 = p_a q_c - f_{ab} + f_{bc}.$$

Under our assumptions, $f_{ab} \approx f_{bc} \approx p_c$ for small-to-moderate $t$ to a reasonable approximation, and therefore $M_1 \approx q_a p_c$, $M_2 \approx p_a q_c$. We then may write

$$\begin{aligned}\max^* D_{ac} &= \min(q_a p_c, p_a q_c, M_1, M_2) \\ &\approx \min(q_a p_c, p_a q_c) = \max D_{ac}.\end{aligned}$$

Along with $D_{ac} < \max D_{ac}$, this implies

$$\max D_{ac} \min^* D_{ac} > D_{ac} \times (\min^* D_{ac} + \max D_{ac} - \max^* D_{ac}),$$

which is algebraically equivalent to

$$\frac{D_{ac}}{\max D_{ac}} - \frac{D_{ac} - \min^* D_{ac}}{\max^* D_{ac} - \min^* D_{ac}} > 0,$$

or $\delta_{a(b)c} > 0$. Putting the above together, we have shown that $\delta_{a(b)c} \approx 0$ for $t = 0$ and $t$ small, but $\delta_{a(b)c} > 0$ for $t$ moderate.

Using very similar arguments, it is possible to show that $\delta_{ab(c)} \geq 0$ during the same time interval, so that the same general mechanisms give the "correct" signal at the C locus. The constrasting result $\delta_{(a)bc} \approx 0$ can be obtained using the same detailed arguments, or more easily can be obtained by noting that $D_{bc}$ remains very close to max $D_{bc}$ during the time interval of interest. Taken together, these arguments suggest that under our assumptions the CDV criteria could indicate selection at either the B or C loci, but not at the A locus.

*A selective sweep model:* A second basic model may be handled without doing any further analysis. For the selective sweep case, we assume $s_c \leq 0$ and $s_C > 0$, so that the selected mutant $c$ will be fixed, but the remaining assumptions are the same. The transient dynamics of allele and haplotype frequencies are the same as in the overdominance model, with perhaps minor differences in time scale; the main difference is in the endpoint of the selection process. Maynard-Smith and Haigh (1974) showed that an allele at a polymorphic locus closely linked to a new favored mutant may readily fix with the selected mutant. In our model, if the $c$ allele fixes in a small number of generations, the time during which $D_{ac}$ and $D_{bc}$ are positive will be very short, since these equal zero once the C locus has become monomorphic. If there has been very little recombination with $A$- or $B$-bearing haplotypes by the time $c$ fixes, $D_{ab}$ will also depart only transiently from zero, because $a$ and $b$ will then nearly fix with $c$. There appears to be only a small time frame in which we could observe any disequilibrium, hence any change in δ values, in the sweep model. The basic reasoning of the previous section again suggests that during this time, a signal for apparent selection is possible from either the selected locus or a nearby neutral locus carrying a rare allele.

**Stochastic simulations:** We have calculated δ values in simulated random samples from a stochastic, neutral diallelic model, to informally investigate the "type I" error in the CDV method. The three-locus neutral model is perhaps the simplest null-model that would be considered for data of the type used for CDV. Hudson (1985) investigated the sampling distribution of the pairwise disequilibrium $D$ using a similar approach. Pairwise $D$ have been treated analytically under the neutral model by Hill (1975), Golding (1984), and Hill and Weir (1988).

Distributions of the three δ values in samples of size $n = 100$ are shown in Figure 4 for $4Nc = 10$, 25, and 100. The histograms of Figure 4 show only the univariate (marginal) distributions of δ values and contain no information about the associations within samples of the

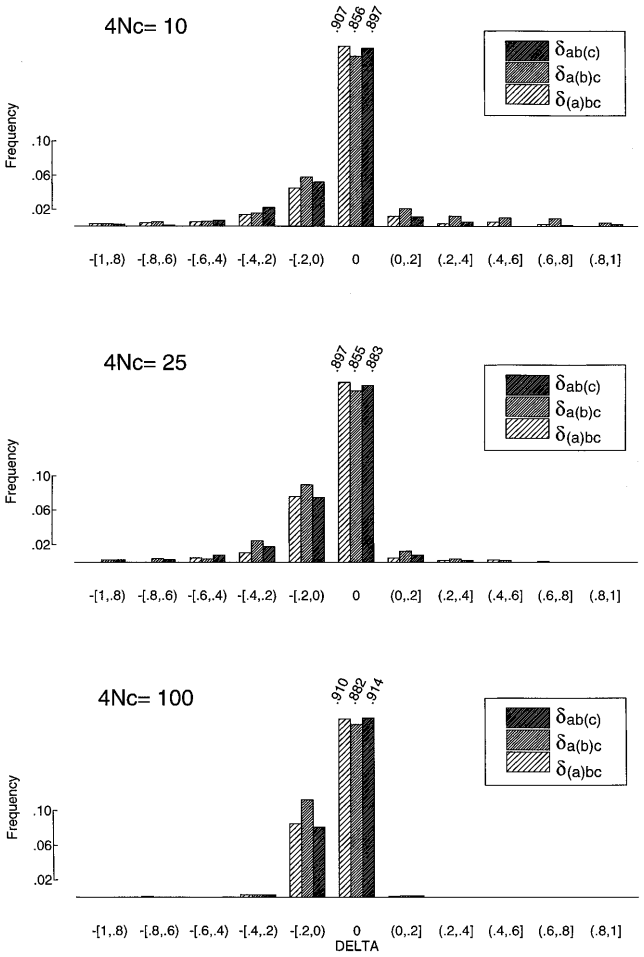


Figure 4.—Distributions of δ values under a stochastic neutral model for three values of $4Nc$ (where $c$ is the recombination rate per generation between the A and C loci and $N$ is the effective population size). Each individual data set consists of $n = 100$ three-locus haplotypes, having exactly two alleles segregating at each locus, with per-locus heterozygosities of at least 0.095. The distributions are based on 1000 independent data sets, generated using a modified program of Hudson (1983, 1985). Bars above $\delta = 0$ are not drawn to scale with the remaining bars; instead the frequencies at $\delta = 0$ are indicated in the figure.

three δ values. At each value of $4Nc$, the $\delta = 0$ class is by far the most common for each of $\delta_{(a)bc}$, $\delta_{a(b)c}$, and $\delta_{ab(c)}$, with $\delta \geq 0$ relatively uncommon. Negative values of δ are more common than positive values when δ departs from zero. Relative to $\delta_{(a)bc}$ and $\delta_{ab(c)}$, $\delta_{a(b)c}$ is more often different from zero.

The frequencies of apparent "hitchhiking" events, obtained by applying the CDV criteria to the samples in Figure 4, are shown in Table 3. For $4Nc = 10$, each locus satisfies the criteria for selection in a small percentage of cases: here one can expect to find a signal for selection at some locus perhaps 8 to 9% of the time, using the CDV criteria in a neutral sample. With $4Nc \geq 25$, however, any apparent signal for "selection" based on the CDV criteria would be unusual. In concordance with

**TABLE 3**

**Frequency of signal for apparent selection: neutral samples ($n = 100$)**

| | Locus | | | |
|---|---|---|---|---|
| | A | B | C | Total |
| $4Nc = 10$ | 0.02 | 0.051 | 0.014 | 0.085 |
| $4Nc = 25$ | 0.009 | 0.016 | 0.006 | 0.031 |
| $4Nc = 100$ | 0.0 | 0.002 | 0.002 | 0.004 |

Frequency of signals for apparent selection at each locus using the CDV criteria for the neutral diallelic samples of Figure 4. Because only one locus per sample may satisfy the CDV criteria for selection, we have added the frequencies of per-locus signals to obtain the totals.

the deterministic simulations, although here there is no selection, we obtain false signals for selection at the B locus more often than at A or C (as expected, the A and C loci give similar results). We take this as further evidence of a "position" effect that favors the middle locus.

**Marker haplotypes from human chromosome *6*:** To illustrate one use of the CDV method, we have calculated δ values in a series of three-locus microsatellite haplotypes in the 6p21.3-22.1 region of human chromosome *6* (see Figure 5). We do not presume any of these marker loci are selected, but suppose instead that perhaps one or more markers could be closely linked to a selected gene.

We used a "sliding window" approach, examining in turn each of the five groups of three adjacent markers among the seven markers shown in Figure 5. Human leukocyte antigen (HLA)-F3′ and myelin oligodendrocyte glycoprotein (MOG)c are dinucleotide repeats closely linked to the HLA F locus and the MOG locus, respectively. HLA-A, a major histocompatibility complex class I locus, is located between the D6S265 and HLA-F3′ markers shown in Figure 5 (Lauer *et al.* 1997; Mosser *et al.* 1997). This region contains other loci of biological and evolutionary interest and has been the focus of recent intensive efforts to map the hereditary hemochromatosis locus, now known to be ~2.2 Mb telomeric to D6S464 (Feder *et al.* 1996; Lauer *et al.* 1997; Mosser *et al.* 1997). The data we used are from a sample of 70 randomly ascertained ethnic Germans and were generously provided by L. Calandro and G. F. Sensabaugh (see Sensabaugh *et al.* 1996). We used an expectation-maximization (EM) algorithm to estimate haplotype frequencies from multilocus genotypes (Baur and Danilovs 1980), working separately with each group of three adjacent markers. The EM algorithm provides haplotype frequency estimates for all possible combinations of alleles at the three loci, some of which have very low estimates and are unlikely to actually be in the sample. For further calculations, we retained only those three-

locus haplotypes in which the constituent two-locus estimates were at least 0.05. Seventeen three-locus haplotypes in all met this minimum frequency threshold: 3 haplotypes of the D6S265/HLA-F3′/MOGc loci, 5 haplotypes of the HLA-F3′/MOGc/D6S258 loci, 3 haplotypes of the MOGc/D6S258/D6S306 loci, 4 haplotypes of the D6S258/D6S306/D6S105 loci, and 2 haplotypes of the D6S306/D6S105/D6S464 loci. All are combinations of the few most common alleles at each locus. We calculated δ values in each of these 17 haplotypes, converting to dialleles by combining the alleles not under consideration into a single class. All 3 haplotypes of the D6S265/HLA-F3′/MOGc loci had disequilibrium patterns conforming to criteria 1 or 2 (Table 4), but none of the remaining 14 haplotypes met these criteria.

## DISCUSSION

**Inferences with CDV:** In the deterministic runs, where the new selected mutant appeared at a terminal locus (the C locus), the CDV method did not distinguish well between the selected locus and a neutral neighbor, especially when a relatively rare allele of the neutral locus was initially linked with the selected mutant. In this case a signal for apparent selection could either be detected from the selected locus or the neutral locus. This situation is unfortunate and somewhat paradoxical, because we have argued that selected mutants that form on rare haplotypes create the most significant linkage disequilibrium in a hitchhiking scenario. To some extent, the CDV method is sensitive to each of the parameters of the model, but we discovered in particular a sensitivity to allele frequencies at the middle locus (the B locus). We showed, using an analytical approach under the assumption of strong selection and tight linkage, that a rare neutral allele at the B locus may easily be mistaken by the CDV criteria for the selected mutant *c*.

Lewontin (1988) showed that the normalized pairwise measure $D'_{ab}$ and other related measures are not in any general sense independent of the underlying allele frequencies $p_a$ and $p_b$, although they are routinely treated as such. The CDV method uses both $D'$ and $D''$ at each pair of the three-locus system, where $D''$ incorporates further one- and two-locus frequency constraints. In light of Lewontin's (1988) results, it is not unexpected that CDV shows a sensitive dependence on allele and haplotype frequencies, as well as on other parameters of the model.

In our deterministic simulations, when the middle locus (the B locus) had the new selected mutant, the CDV method gave correct inferences in a large majority of runs. It is difficult to put this attractive result into practice in the inference setting, because a signal for apparent selection at the B locus could indeed reflect selection at the locus or could be a false signal of the type that was commonly observed when *c* was the selected allele. One remedy might be to confine inferences to terminal loci, perhaps obtaining additional markers that could place any locus of interest at the "A" or "C" positions of our model. This assumes we could be virtually certain about inferences at terminal loci, an assertion that is contradicted by the fraction of deterministic runs of Tables 1 and 2 in which a signal appears at an unselected terminal locus. It further seems possible that a generalization of our analytical approach, which relaxes assumptions about position, could show that inferences about terminal loci may not be reliable in the presence of rare neutral alleles. We think at present that the CDV method may not allow for high-precision

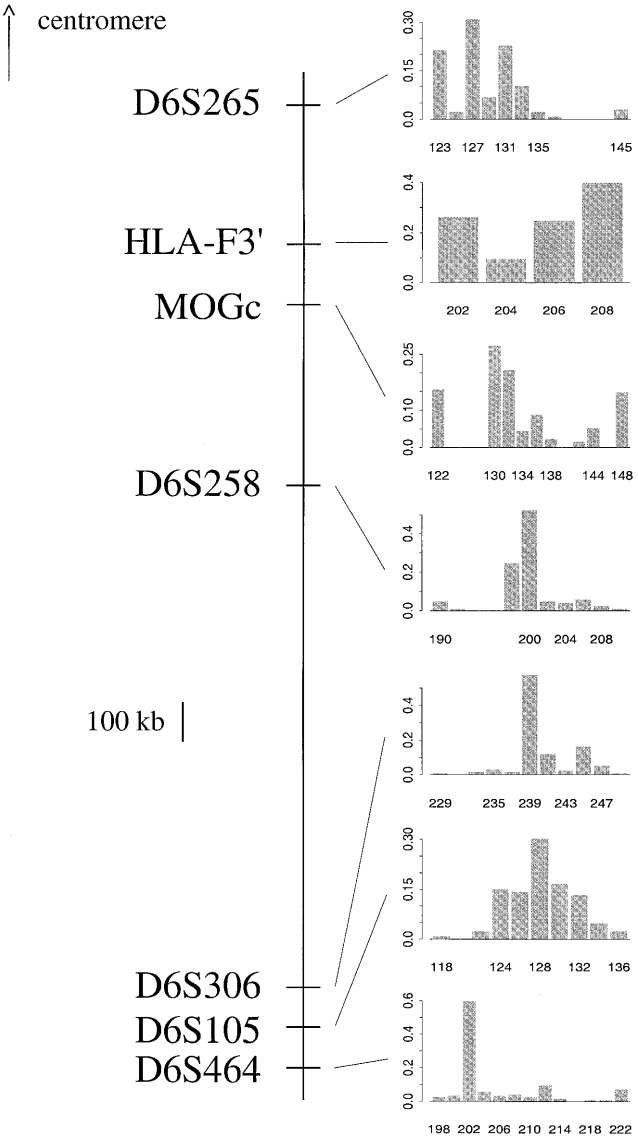


Figure 5.—Physical map of seven dinucleotide repeat markers in the 6p21.3-22.1 region of human chromosome *6*. Approximate intermarker distances are based on the YAC contig and STS maps of Mosser *et al.* (1997). Allele frequency distributions are for the sample of 70 ethnic Germans provided by L. Calandro and G. F. Sensabaugh (Sensabaugh *et al.* 1996). The *x* and *y* axes of the histograms are labeled according to repeat number and frequency in the sample, respectively.

**TABLE 4**

**Ethnic German sample haplotypes with signal for selection**

| Haplotype | | | $D$ | $D'$ | $D''$ | δ |
|---|---|---|---|---|---|---|
| (D6S265-123) | HLAF3′-206 | MOGc-130 | 0.123 | 0.676 | 0.626 | 0.050 |
| D6S265-123 | (HLAF3′-206) | MOGc-130 | 0.103 | 0.661 | 0.405 | <u>0.256</u> |
| D6S265-123 | HLAF3′-206 | (MOGc-130) | 0.131 | 0.816 | 0.780 | 0.036 |
| (D6S265-131) | HLAF3′-208 | MOGc-132 | 0.098 | 0.767 | 0.665 | 0.102 |
| D6S265-131 | (HLAF3′-208) | MOGc-132 | 0.122 | 0.767 | 0.767 | 0.000 |
| D6S265-131 | HLAF3′-208 | (MOGc-132) | 0.089 | 0.635 | 0.420 | <u>0.215</u> |
| (D6S265-127) | HLAF3′-202 | MOGc-148 | 0.099 | 0.929 | 0.929 | 0.000 |
| D6S265-127 | (HLAF3′-202) | MOGc-148 | 0.079 | 0.781 | 0.698 | <u>0.083</u> |
| D6S265-127 | HLAF3′-202 | (MOGc-148) | 0.122 | 0.648 | 0.633 | 0.015 |

Three-locus haplotypes of the ethnic German sample ($2n = 140$) that satisfy criteria 1 or 2. Pairwise $D$ and $D'$ are given for the two-locus haplotype in each row composed of the alleles not in parentheses. $D''$ is calculated for the same haplotype under the constraints imposed by the allele in parentheses. δ is the difference $|D'| - |D''|$. The δ for each haplotype that indicates possible hitchhiking, according to the CDV method, is underlined.

inferences about the location of selected mutants; on this point we depart from Robinson *et al.* (1991a).

The stochastic simulations showed that patterns of linkage disequilibrium conforming to criteria 1 or 2 are uncommon for $4Nc \geq 10$ and highly unusual for $4Nc$ as large as 100. Here, we think there is potential inference value in the CDV method, because a simple neutral model can apparently be ruled out if either criteria 1 or 2 is met in a moderate-sized sample, with $4Nc$ on the order of 100. At this point, other nonselective alternative hypotheses (such as the neutral model with population structure or migration) cannot immediately be ruled out; this requires work beyond our current scope.

Although we do not think that CDV can very accurately distinguish the particular locus that has the selected allele, we do think that CDV can be used to screen for fairly localized regions that may have a recent history of hitchhiking (in general agreement with Robinson *et al.* 1991a). The basic requirements appear to be that the terminal loci span at least a distance of $4Nc = 10$ (with the third locus roughly intermediate), that there is a standard minimum level of heterozygosity $H \geq 0.095$ at each locus, and that there is moderately strong, but not complete, linkage disequilibrium in the region.

**Selected mutations and linked markers at equilibrium:** We now describe a simple model of recurrent selected mutations and address some implications for CDV and similar methods. The simplest model assumes that selected alleles arise at random points in the genome. If such events are rare, the influence of new selected alleles on linked loci is transient: eventually the new mutant reaches equilibrium, and recombination, mutation, and genetic drift again dominate the dynamics of linked loci. Under this simple model, neutral alleles linked to a new overdominant mutant will increase in frequency and may reach high levels of disequilibrium, but do not generally fix (because the overdominance mode tends to preserve extant variation). Two such loci will return to neutral frequency and phase equilibria, respectively, at rates $1 - 1/2N$, the rate of loss of heterozygosity at either locus (with $N$ the effective population size; see, *e.g.*, Crow and Kimura 1970), and $1 - c - 1/2N$, the rate of decay of linkage disequilibrium between the loci (for the random union of gametes model, where $c$ is the recombination rate; Hill 1974). In the selective sweep case, if a neutral allele fixes with the new mutant, the time until polymorphism could be reestablished at the neutral locus is on the order of $1/\mu$, where $\mu$ is the neutral mutation rate (Crow and Kimura 1970). If either overdominant or favored mutants reoccur in a particular region over relatively short time scales, and the recovery of linkage equilibrium or polymorphism is inadequate, reperturbation by successive hitchhiking events may not be detectable. Even if selected mutants appear only rarely, the availability of adequate polymorphism at closely linked sites, on which linkage disequilibrium could be recorded, may be in question; for if $\theta = 4N\mu$ is small, a majority of linked neutral sites will be monomorphic. We have further claimed that disequilibrium created by hitchhiking is primarily connected to rare events in which selected mutants appear on low-frequency haplotypes. In particular, these impediments suggest that in chromosomal regions thought to be subject to recurrent selective sweeps (Aguadé *et al.* 1989; Begun and Aquadro 1994, 1995), the linkage disequilibrium that is indeed observed is primarily the result of mutational or other events that occurred since the most recent sweep.

For tightly linked loci, patterns of linkage disequilibrium conforming to criteria 1 or 2 persist approximately as long as the time required for the new mutant to reach equilibrium (Thomson 1977, and our deterministic simulations). If we assume strong selection and large $N$, and confine our attention only to mutants that invade

the population, the expected time until the new mutant fixes in the selective sweep model is approximately

$$\int_{1/2N}^{1-1/2N} 2/sx(1-x)\,dx \approx \frac{4}{s}\ln(2N)$$

generations (Ewens 1973, 1979, p. 149). Here, $sx(1-x)/2$ is the approximate deterministic change per generation in the frequency of a favored allele $a$, where fitnesses are $w_{AA} = 1 - s/2$, $w_{Aa} = 1$, $w_{aa} = 1 + s/2$. Using the same reasoning in the symmetric overdominance model, with fitnesses $w_{AA} = 1 - s$, $w_{Aa} = 1$, $w_{aa} = 1 - s$ (so that the fitness differential between the most extreme genotypes is $s$ in both cases), the expected time for the new mutant to reach the interior polymorphism is approximately

$$\int_{1/2N}^{1/2-1/2N} 1/sx(1-x)(1-2x)\,dx \approx \frac{3}{s}\ln(2^{-1/3}N)$$

generations. These persistence times can be small relative to the times required for the recovery of linkage equilibrium or neutral levels of polymorphism. For example, if $N$ is $10^5$, $s = 0.01$, and $\mu = 10^{-5}$, the persistence time for CDV-type patterns of linkage disequilibrium is $<5000$ generations in the selective sweep model, whereas if most extant variation is lost during the sweep, $10^5$ generations on average are required to reestablish polymorphism at monomorphic sites, and during this period no new CDV-type patterns could be observed.

**Human chromosome *6* haplotypes:** In Table 4, we showed three haplotypes of the D6S265/HLA-F3′/MOGc loci that met the CDV criteria for hitchhiking. HLA-F3′ and MOGc are physically close, so we must make a rough assessment of $4Nc$ between these loci if we wish to compare the data with the neutral simulations of Figure 4 and Table 3. Although there is apparently no family data that give precise estimates of the recombination fraction between HLA-F3′ and MOGc, the physical distance between these loci is known to be approximately 100–150 kb, based on YAC contig and STS maps (Mosser *et al.* 1997; Human Genome Data Base 1997). For estimation purposes, we will assume the distance is 100 kb and use $N = 2000$, perhaps a conservatively low value for a modern European population. If we use the crude conversion 1 Mb ≈ 1.16 cM [obtained by observing that the genome size is equivalently 3200 Mb or 3702 cM in human females (The Human Transcript Map 1996)], we conclude that $4Nc$ between HLA-F3′ and MOGc is ∼9. Thus, the D6S265/HLA-F3′/MOGc haplotype appears to span a distance over which criteria 1 or 2 are not commonly met in the simple neutral model. The setting here is not directly analogous to the null-model calculations of Figure 4 and Table 3 for two main reasons: (i) different three-locus marker haplotypes may share an allele at one or more loci, introducing dependencies not present in the simulated neutral haplotypes; (ii) it is well known that the "infinite alleles" mutation model used for the neutral simulations does not apply to microsatellite loci (see, *e.g.*, Valdes *et al.* 1993). However, the role of the mutation model, especially given the time scale for mutational events relative to the duration of hitchhiking events, should be minor. Finally, if the scaling of $4Nc$ is approximately correct, the chances under the neutral model that even one of the D6S265/HLA-F3′/MOGc haplotypes would meet the CDV criteria appear to be small.

We conclude that hitchhiking with one or more selected alleles, closely linked to the D6S265/HLA-F3′/MOGc loci, is a plausible explanation for the patterns of linkage disequilibrium observed in these haplotypes. Three apparently distinct haplotypes meet criteria 1 or 2, suggesting that hitchhiking with overdominant alleles is the more likely scenario: the data would seem to require otherwise that several favored alleles in the region are simultaneously being selected for, or that an ancestral haplotype bearing a favored allele has experienced several mutation events. We have also argued that the loss of variation under the selective sweep model poses a serious problem for observing disequilibrium, making it unlikely that disequilibrium created specifically by selectively favored alleles would ever be observed. While we have scaled back previous efforts to infer the precise location at which selection has acted, our results are consistent with other work on selection in this region of the human genome (Klitz and Thomson 1987; Satta *et al.* 1994; Parham and Ohta 1996). Our main intention in this example is to demonstrate that evidence for historical selection processes may indeed be found in the patterns of linkage disequilibrium we have focused on, in our investigation of the CDV method.

We thank C. H. Langley, who read an earlier draft of the manuscript and made suggestions that led to substantial revisions. The human chromosome *6* haplotypes were collected by L. Calandro and G. F. Sensabaugh, who generously allowed us to use them here. We thank D. Cutler and A. D. Long for discussion and suggestions. An anonymous reviewer made suggestions that improved the presentation. This work was supported by National Institutes of Health grants HD-12731, GM-56688, and 5 T32 GM-07127.

## 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.

Baur, M. P., and J. A. Danilovs, 1980 Population analysis of HLA-A, B, C and DR and other genetics markers, pp. 955–993 in *Histocompatibility Testing 1980*, edited by P. Terasaki. University of California, Tissue Typing Laboratory, Los Angeles.

Begun, D. J., and C. F. Aquadro, 1994 Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of *Drosophila*: selection and geographic differentiation. Genetics **136:** 155–171.

Begun, D. J., and C. F. Aquadro, 1995 Evolution at the tip and base of the X chromosome in an African population of *Drosophila melanogaster*. Mol. Biol. Evol. **12:** 382–390.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Burgess Publishing Co., Minneapolis.

Ewens, W. J., 1973 Conditional diffusion processes in population genetics. Theor. Pop. Biol. **4:** 21–30.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.
Feder, J. N., A. Gnirke, W. Thomas, Z. Tsuchihashi, D. A. Ruddy *et al.*, 1996 A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat. Genet. **13:** 399–408.
Feldman, M. W., I. Franklin and G. J. Thomson, 1974 Selection in complex genetic systems I. The symmetric equilibria of the three-locus symmetric viability model. Genetics **76:** 135–162.
Geiringer, H., 1944 On the probability theory of linkage in Mendelian heredity. Ann. Math. Stat. **15:** 25–57.
Golding, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108:** 257–274.
Grote, M. N., 1996 Models of genetic selection and the Human Leukocyte Antigen loci. Ph.D. Thesis, University of California, Berkeley.
Hartl, D. L., and A. G. Clark, 1989 *Principles of Population Genetics*. Sinauer Associates, Inc., Sunderland, MA.
Hedrick, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics **117:** 331–341.
Hill, W. G., 1974 Disequilibrium among several linked neutral genes in finite population I. Mean changes in disequilibrium. Theor. Pop. Biol. **5:** 366–392.
Hill, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite populations. Theor. Pop. Biol. **8:** 117–126.
Hill, W. G., and B. S. Weir, 1988 Variances and covariances of squared linkage disequilibria in finite populations. Theor. Pop. Biol. **33:** 54–78.
Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. **23:** 183–201.
Hudson, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109:** 611–631.
Human Genome Data Base, 1997 http://www.gdb.org
The Human Transcript Map, 1996 http://www.ncbi.nlm.nih.gov/SCIENCE96
Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.
Klitz, W., and G. Thomson, 1987 Disequilibrium pattern analysis. II. Application to Danish HLA-A and B locus data. Genetics **116:** 633–643.
Lauer, P., N. C. Meyer, C. E. Prass, S. M. Starnes, R. K. Wolff *et al.*, 1997 Clone-contig and STS maps of the hereditary hemochromatosis region on human chromosome 6p21.3-p22. Genome Res. **7:** 457–470.
Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.
Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics **120:** 849–852.
Maynard-Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.
Mosser, J., A. M. Jouanolle, G. Gandon, N. Andrieux, A. Hampe *et al.*, 1997 A YAC contig and an STS map spanning at least 3.9 megabasepairs telomeric to HLA-A. Immunogenet. **45:** 447–451.
Ohta, T., and M. Kimura, 1975 The effect of a selected linked locus on heterozygosity of neutral alleles (the hitchhiking effect). Genet. Res. **25:** 313–325.
Parham, P., and T. Ohta, 1996 Population biology of antigen presentation by MHC class-I molecules. Science **272:** 67–74.
Robinson, W. P., A. Cambon-Thomsen, N. Borot, W. Klitz and G. Thomson, 1991a Selection, hitchhiking and disequilibrium analysis at three linked loci with application to HLA data. Genetics **129:** 931–948.
Robinson, W. P., M. A. Asmussen and G. Thomson, 1991b Three-locus systems impose additional constraints on pairwise disequilibria. Genetics **129:** 925–930.
Satta, Y., C. O'huigen, N. Takahata and J. Klein, 1994 Intensity of natural selection at the major histocompatibility complex loci. Proc. Natl. Acad. Sci. USA **91:** 7184–7188.
Sensabaugh, G. F., L. Calandro, T. Thorsen, L. Barcellos, J. Griggs *et al.*, 1996 Commentary. Blood Cells Mol. Dis. **22:** 194a–194b.
Stephan, W., and C. H. Langley, 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermillion* and *forked* loci. Genetics **121:** 89–99.
Thomson, G., 1977 The effect of a selected locus on linked neutral loci. Genetics **85:** 753–788.
Thomson, G., and M. P. Baur, 1984 Third order linkage disequilibrium. Tissue Antigens **24:** 250–255.
Valdes, A. M., M. Slatkin and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

Communicating editor: G. B. Golding

## APPENDIX: THREE-LOCUS DETERMINISTIC RECURSIONS

In the three-locus diallelic model, the eight haplotypes (gametes) are *ABC*, *ABc*, *AbC*, *Abc*, *aBC*, *aBc*, *abC*, *abc*, and their respective frequencies in a given generation are $x_1, \ldots x_8$, $\sum_{i=1}^{8} x_i = 1$. Let

$$\bar{w}_i = \sum_{j=1}^{8} w_{ij}x_j \quad i = 1, \ldots 8,$$

where $w_{ij}$ is the fitness of the genotype formed from haplotypes $i$ and $j$, and let $\bar{w} = \sum_{i=1}^{8} \bar{w}_i x_i$. After selection and recombination, the haplotype frequencies are given by

$$\begin{aligned}
\bar{w}x_1' &= \bar{w}_1x_1 - r_1\alpha_1 - r_2\beta_1 + r_1r_2\gamma_1\\
\bar{w}x_2' &= \bar{w}_2x_2 + r_1\alpha_2 + r_2\beta_1 - r_1r_2\gamma_1\\
\bar{w}x_3' &= \bar{w}_3x_3 + r_1\alpha_3 + r_2\beta_2 + r_1r_2\gamma_2\\
\bar{w}x_4' &= \bar{w}_4x_4 + r_1\alpha_4 - r_2\beta_2 - r_1r_2\gamma_2\\
\bar{w}x_5' &= \bar{w}_5x_5 + r_1\alpha_1 + r_2\beta_3 - r_1r_2\gamma_1\\
\bar{w}x_6' &= \bar{w}_6x_6 - r_1\alpha_2 - r_2\beta_3 + r_1r_2\gamma_1\\
\bar{w}x_7' &= \bar{w}_7x_7 - r_1\alpha_3 + r_2\beta_4 - r_1r_2\gamma_2\\
\bar{w}x_8' &= \bar{w}_8x_8 - r_1\alpha_4 - r_2\beta_4 + r_1r_2\gamma_2,
\end{aligned}$$

where

$$\begin{aligned}
\alpha_1 &= w_{16}x_1x_6 + w_{17}x_1x_7 + w_{18}x_1x_8 - w_{25}x_2x_5 - w_{35}x_3x_5 - w_{45}x_4x_5\\
\alpha_2 &= w_{16}x_1x_6 - w_{25}x_2x_5 - w_{27}x_2x_7 - w_{28}x_2x_8 + w_{36}x_3x_6 + w_{46}x_4x_6\\
\alpha_3 &= w_{17}x_1x_7 + w_{27}x_2x_7 - w_{35}x_3x_5 - w_{36}x_3x_6 - w_{38}x_3x_8 + w_{47}x_4x_7\\
\alpha_4 &= w_{18}x_1x_8 + w_{28}x_2x_8 + w_{38}x_3x_8 - w_{45}x_4x_5 - w_{46}x_4x_6 - w_{47}x_4x_7\\
\beta_1 &= w_{14}x_1x_4 + w_{16}x_1x_6 + w_{18}x_1x_8 - w_{23}x_2x_3 - w_{25}x_2x_5 - w_{27}x_2x_7\\
\beta_2 &= w_{14}x_1x_4 - w_{23}x_2x_3 - w_{36}x_3x_6 - w_{38}x_3x_8 + w_{45}x_4x_5 + w_{47}x_4x_7\\
\beta_3 &= w_{16}x_1x_6 - w_{25}x_2x_5 + w_{36}x_3x_6 - w_{45}x_4x_5 - w_{58}x_5x_8 + w_{67}x_6x_7\\
\beta_4 &= w_{18}x_1x_8 - w_{27}x_2x_7 + w_{38}x_3x_8 - w_{47}x_4x_7 + w_{58}x_5x_8 - w_{67}x_6x_7\\
\gamma_1 &= 2w_{16}x_1x_6 + w_{18}x_1x_8 - 2w_{25}x_2x_5 - w_{27}x_2x_7 + w_{36}x_3x_6 - w_{45}x_4x_5\\
\gamma_2 &= w_{18}x_1x_8 - w_{27}x_2x_7 + w_{36}x_3x_6 + 2w_{38}x_3x_8 - w_{45}x_4x_5 - 2w_{47}x_4x_7.
\end{aligned}$$

To complete one generation, we need only introduce mutation, which is unidirectional from *a*, *b*, and *c* to *A*, *B*, and *C*, respectively, all at rate $\mu$ per generation. After mutation, the haplotype frequencies are

$$\begin{aligned}
x_1'' &= x_1' + \mu(x_2' + x_3' + x_5') + \mu^2(x_4' + x_6' + x_7') + \mu^3x_8'\\
x_2'' &= (1-\mu)x_2' + \mu(1-\mu)(x_4' + x_6') + \mu^2(1-\mu)x_8'\\
x_3'' &= (1-\mu)x_3' + \mu(1-\mu)(x_4' + x_7') + \mu^2(1-\mu)x_8'\\
x_4'' &= (1-\mu)^2x_4' + \mu(1-\mu)^2x_8'\\
x_5'' &= (1-\mu)x_5' + \mu(1-\mu)(x_6' + x_7') + \mu^2(1-\mu)x_8'\\
x_6'' &= (1-\mu)^2x_6' + \mu(1-\mu)^2x_8'\\
x_7'' &= (1-\mu)^2x_7' + \mu(1-\mu)^2x_8'\\
x_8'' &= (1-\mu)^3x_8'.
\end{aligned}$$

This completes one generation of the recursion.