Copyright © 1987 by the Genetics Society of America

Disequilibrium Pattern Analysis. I. Theory

Glenys Thomson and William Klitz

Department of Genetics, University of California, Berkeley, California 94720

Manuscript received September 3, 1986
Revised copy accepted May 2, 1987

ABSTRACT

We have developed a method, disequilibrium pattern analysis, for examining the disequilibrium distribution of the entire array of two locus multiallelic haplotypes in a population. It is shown that a selected haplotype will produce a distinct pattern of linkage disequilibrium values for all generations while the selection is acting. This pattern will also presumably be maintained for many generations after the selection event, until the disequilibrium pattern is eventually broken down by genetic drift and recombination. Related haplotypes, sharing an allele with a selected haplotype, assume a value of linkage disequilibrium proportional to the frequency of the unshared allele and have a single negative value of the normalized linkage disequilibrium. The analysis assumes zero linkage disequilibrium for all allelic combinations initially. The same basic results continue to apply if the selection involves a new mutant, the occurrence of which creates linkage disequilibrium for some haplotypes. The disequilibrium pattern predicted under selection is robust with respect to the influence of migration and random genetic drift. This method is applicable to population data having linked polymorphic loci including that determined from protein or DNA sequencing.

THE linkage disequilibrium of haplotypes (chromosome or gametic types) formed from tightly linked loci may indicate the nature of past evolutionary events, particularly selection events, acting on the genome. Unlike genotypic ratios, which assume Hardy-Weinberg proportions in one generation in random mating populations, disequilibrium present in haplotypes is retained over generations and its decay over time is a function of recombination between loci. Tightly linked polymorphic loci will retain evidence of distorted haplotypic frequencies for many generations. Thus, the patterns and extent of linkage disequilibrium potentially provide more information for differentiating selection from other evolutionary events than does examination of single locus genetic variation.

In this paper we describe a method, disequilibrium pattern analysis, for examining the disequilibrium distribution of the entire array of haplotypes in a polymorphic two locus system. This method identifies the constraints on the disequilibrium values that a haplotype can assume, and the distribution of the entire array of haplotypes in the disequilibrium space. It is shown that selection leads to specific quantifiable disequilibrium patterns that are robust with respect to the influence of migration and random genetic drift.

Our impetus for developing this methodology was to study the population genetics and evolutionary history of the major histocompatibility complex in humans (HLA). The HLA system contains a number of closely linked and highly polymorphic loci, which have strong linkage disequilibrium. There exists an extensive amount of data on the genetic variation for HLA genes in different populations (for example, see the *Histocompatibility Testing* volumes for 1972, 1980 and 1984, DAUSSET and COLOMBANI 1973; TERASAKI 1980 and ALBERT, BAUR and MAYR 1984).

In this paper we discuss the theoretical aspects of disequilibrium pattern analysis for detecting evolutionary forces acting on a population. The companion paper (KLITZ and THOMSON 1987) applies disequilibrium pattern analysis to HLA *A* and *B* locus data from a large Danish study (HANSEN *et al.* 1979).

RESULTS

Defining the disequilibrium space: The linkage disequilibrium parameter D_{ij} is defined as

$$D_{ij} = x_{ij} - p_{A_i} q_{B_j} \tag{1}$$

where x_{ij} is the observed frequency of the haplotype A_iB_j, p_{A_i} is the frequency of the allele A_i and q_{B_j} that of B_j, so that the linkage disequilibrium is the difference between the observed haplotype frequency x_{ij} and $p_{A_i}q_{B_j}$, the haplotype frequency expected if the alleles are associated at random. The disequilibrium values of the haplotypes of a population are constrained by the relationships

$$\sum_{i=1}^{r} D_{ij} = 0 = \sum_{j=1}^{s} D_{ij} \tag{2}$$

where r is the number of alleles at the A locus, and s the number of alleles at the B locus.

This constraint is most simply illustrated in the case of a two locus, two allele (A_1, A_2 and B_1, B_2) model, where the four haplotype frequencies can be expressed as

$$f(A_1B_1) = p_{A_1}q_{B_1} + D, \qquad f(A_2B_2) = p_{A_2}q_{B_2} + D \tag{3a}$$

$$f(A_1B_2) = p_{A_1}q_{B_2} - D, \qquad f(A_2B_1) = p_{A_2}q_{B_1} - D. \tag{3b}$$

Note that $p_{A_1} + p_{A_2} = 1$ etc., and D is the measure of linkage disequilibrium ($= f(A_1B_1)f(A_2B_2) - f(A_1B_2)f(A_2B_1)$). If the alleles A_1 and B_1 occur together in the haplotype A_1B_1 more often than expected under random association, that is, $D > 0$, then the haplotype A_2B_2 is also in the positive disequilibrium space. Although the frequency of the haplotype A_2B_2 will be different from that of the haplotype A_1B_1, unless $p_{A_1} = q_{B_1} = 0.5$, the constraints given in (2), and illustrated in (3a), imply that these two haplotypes will have an equal value of linkage disequilibrium. The positive disequilibria are balanced in the negative space by the two haplotypes A_1B_2 and A_2B_1, which each share one allele in common with the haplotypes in the positive space. This well-known representation, given in equations (3a) and (3b), can be easily extended to the two locus multi-allelic case.

Considering two locus multi-allelic cases we have shown that when selection has favored a particular haplotype, each of the haplotypes of the entire population can be assigned to one of three classes which reflect the dynamics of inter-locus association under selection: (1) favored haplotypes, which have strong positive disequilibrium values, (2) related haplotypes having negative values, which share just one allele with the favored haplotypes and (3) unrelated haplotypes having weaker positive disequilibria, which share no alleles with the favored haplotypes. (The selection on a haplotype may be due to a hitchhiking event, or to selection specifically on the allelic combination of the haplotype.)

The maximum value that the linkage disequilibrium D can take is a function of the allele frequencies. A normalized or standardized disequilibrium value, denoted D', which ranges in value from -1 to $+1$, is defined as (LEWONTIN 1964)

$$D'_{ij} = D_{ij}/D_{\max} \tag{4a}$$

where

$$D_{\max} = p_i(1 - q_j) \text{ when } D > 0 \text{ and } p_i \leq q_j \tag{4b}$$

$$= (1 - p_i)q_j \text{ when } D > 0 \text{ and } p_i \geq q_j \tag{4c}$$

$$= p_iq_j \text{ when } D < 0 \text{ and } p_i + q_j \leq 1 \tag{4d}$$

$$= (1 - p_i)(1 - q_j) \text{ when } D < 0 \text{ and } p_i + q_j \geq 1. \tag{4e}$$

When there are multiple alleles with fairly even frequencies, such as with HLA data, this last condition, that is, $p_i + q_j \geq 1$, rarely holds (BAUR and GRANGE 1983).

Effect of selection on the disequilibrium space: Theoretical results have been obtained for deterministic models in which the haplotype A_1B_1 increases in frequency due to selection (see APPENDIX). The results apply whether the haplotype A_1B_1 increases in frequency via a hitchhiking event (THOMSON 1977) or via selection for this haplotype. We consider the case where all linkage disequilibrium values are initially zero, as well as the case where one of the alleles A_1 or B_1 of the selected haplotype is a new mutant. We have used iterations of various selection models to evaluate the effect of recombination, and the results we now list apply in all cases.

If initially there is no linkage disequilibrium (*i.e.* disequilibrium between the alleles at our two loci of interest equals zero), then some simple relationships for the linkage disequilibrium values which result from the selection event can be given (see APPENDIX). These basic results hold in every generation while the selection is operating.

1. The positive disequilibrium of the haplotype A_1B_1, which increases in frequency as a result of the hitchhiking event, and the unrelated haplotypes A_iB_j, $i = 2, \ldots, r, j = 2, \ldots, s$, will be exactly balanced by the negative disequilibrium values of the related haplotypes A_iB_1, $i = 2, \ldots, r$ and A_1B_j, $j = 2, \ldots, s$.

2. Further, the negative disequilibria of the A_iB_1 haplotypes, $i = 2, \ldots, r$ will be proportional to the A_i allele frequency, while the normalized disequilibrium values for all these haplotypes will be equal when $p_{A_i} + p_{B_1} \leq 1$.

3. Similarly, for the A_1B_j haplotypes, $j = 2, \ldots, s$, the linkage disequilibria will be proportional to the B_j allele frequency, and the normalized disequilibria values will be equal for all these haplotypes, when $p_{A_1} + p_{B_j} \leq 1$. In most cases this differs from the constant normalized disequilibrium value for the A_iB_1 haplotypes.

4. The normalized disequilibrium value of the A_1B_1 haplotype will equal the larger in absolute value of the normalized A_iB_1 and A_1B_j disequilibrium values.

5. The positive disequilibria of the unrelated haplotypes A_iB_j ($i = 2, \ldots, r, j = 2, \ldots, s$) are proportional to the A_i and B_j frequencies, but the normalized disequilibrium values will often not be equal, depending on the allele frequencies.

These features are illustrated in Figures 1 and 2, which give plots of the linkage disequilibria, and normalized linkage disequilibria values, respectively, from the iterations of a selection model. The model considers three alleles at the A locus, and four alleles at the B locus, giving 12 haplotypes in the population. The vertical axis is $p_{A_i}q_{B_j}$, the expected frequency of

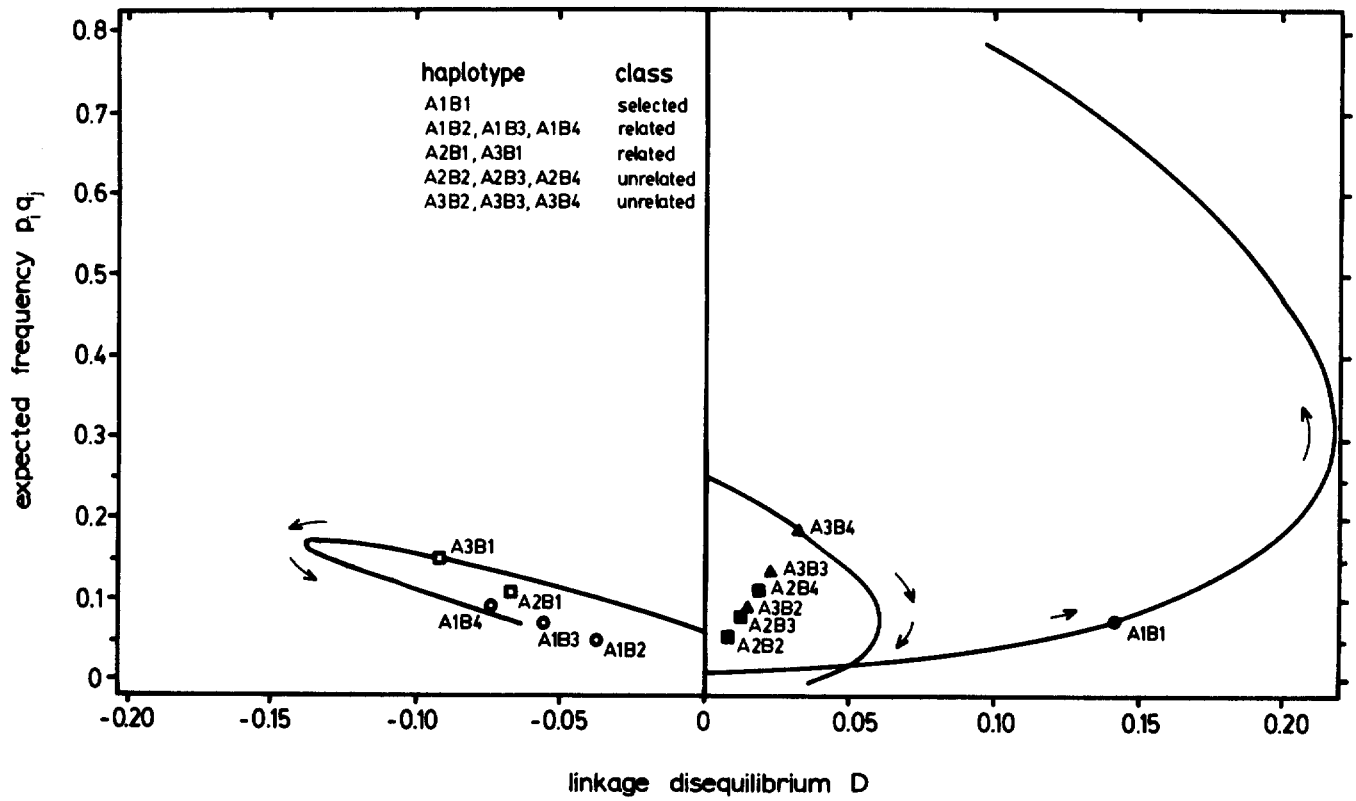

FIGURE 1.—Results of a deterministic two-locus selection model having three and four alleles at each locus, showing the linkage disequilibrium, D, of the twelve haplotypes during directional selection (selection coefficient 0.5) favoring the A_1B_1 haplotype. The initial generation is in linkage equilibrium (all haplotypes at $D = 0$). The trajectory of one haplotype from each class is shown through 25 generations. The location of each haplotype at generation eight is indicated. The initial frequencies for the alleles A_1, A_2 and A_3 are 0.03, 0.37 and 0.60, respectively, and for the alleles B_1, B_2, B_3 and B_4 are 0.10, 0.20, 0.30 and 0.40.

the haplotype if there is no linkage disequilibrium. This choice of axis allows easy discrimination between haplotypes with low and high frequency alleles, and permits easy identification of the proportionality of the negative linkage disequilibrium values with the frequency of the unshared allele. The model is strictly deterministic in that the effects of drift are not considered. Selection favors one haplotype, A_1B_1, and we consider directional selection with a selection coefficient of 0.5 applied to all genotypes lacking the favored haplotype. A large selection coefficient was chosen in order to observe rapid movement of haplotypes in the disequilibrium space: the results are not qualitatively different when other selection values are used. Initial allele frequencies were chosen to differ, the initial disequilibrium values of all haplotypes were zero, and recombination was zero. The favored haplotype was started at a low frequency.

The increase and then decrease in the value of the linkage disequilibrium D for the A_1B_1 haplotype with time (Figure 1) relates to the fact that the maximum value that D can take is a function of the allele frequencies. Note that the normalized linkage disequilibrium D' increases over time (Figure 2).

The linkage disequilibria of each of the classes of related haplotypes, that is (A_1B_2, A_1B_3, and A_1B_4) and (A_2B_1 and A_3B_1), respectively, form a line in the negative space passing through the point $D = 0$, $pq = 0$ in the D space (Figure 1) and each set of related haplotypes assumes a single D value (Figure 2). Unrelated haplotypes (A_2B_2, A_2B_3, A_2B_4 and A_3B_2, A_3B_3, A_3B_4) fall on a line in the D space but form more complicated alignments in the D′ space because the value of D_{max} varies with the relative allele frequencies (see APPENDIX).

These results are also illustrated in case A of Table 1 where the frequencies of five alleles at the A locus and four alleles at the B locus, and the normalized linkage disequilibria (D') of all the pairwise combinations of alleles, are given at generations 0 and 50. All initial disequilibrium values are zero, the selection coefficient of the A_1B_1 haplotype is 0.1, and we assume no recombination. At generation 50 the related haplotypes have D'_{1j} ($j = 2, \ldots, 4$) values which are all equal to -0.334, and D'_{i1} ($i = 2, \ldots, 5$) values which are all equal to -0.905. The normalized disequilibrium of the A_1B_1 haplotype has a value of 0.905 which equals the larger in absolute value of the normalized disequilibria of the related haplotypes. The positive disequilibria of the A_iB_j haplotypes ($i = 2, \ldots, 5$; $j = 2, \ldots, 4$) assume a number of different normalized values, depending on the relative frequencies of the constituent alleles.

As detailed in the APPENDIX, when one of the alleles

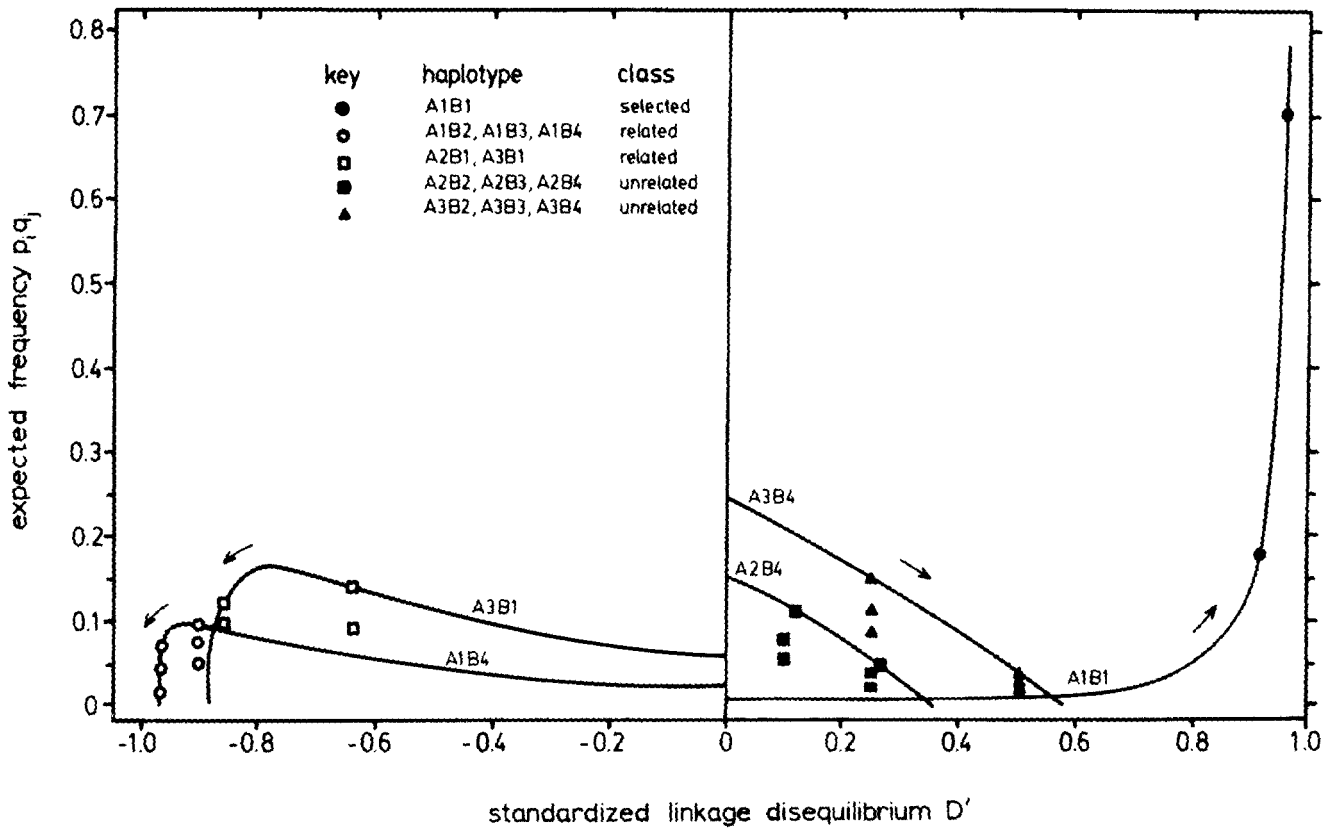

FIGURE 2.—Deterministic selection model identical to that in Figure 1, plotted against standardized linkage disequilibrium D'. The location of haplotypes at the 8th and 16th generations is indicated, with the trajectories over 25 generations shown for the selected haplotype, one from each of the related classes of haplotypes and one from each of the unselected classes.

A_1 or B_1, say B_1, is a new mutant and forms the haplotype A_1B_1 from the haplotype A_1B_2, and the haplotype A_1B_1 is selected, the same basic results outlined above continue to apply, if we assume zero linkage disequilibrium before the new mutant arose. One exception to the general results is that for the related haplotype bearing the allele from which the new mutant arose, $D'_{12}(n)$ will be larger in absolute value than the value of $D'_{1j}(n)$ ($j = 3, \ldots, s$), where n is the number of generations since the new mutation arose (and hence since selection has been operating), (see cases B, C and D in Table 1). The occurrence of the new mutant creates nonzero linkage disequilibrium for a number of other haplotypes (see equations (I-15a)–(I-15e), and cases B, C and D of Table 1). One point of specific interest is that the D'_{i1} ($i = 2, \ldots, r$) values will all equal -1 initially, and will maintain this value if recombination is zero (see case B of Table 1).

The case of nonzero recombination has been considered by computer iteration of various deterministic selection models (see APPENDIX and Table 1). The results obtained for the case of zero recombination all continue to apply with recombination. Two examples where B_1 is a new mutant forming the haplotype A_1B_1 from the haplotype A_1B_2, with recombination $R = 0.01$ and selection coefficient 0.10 for the haplotype A_1B_1 are given in Table 1. In case C, the alleles A_1 and B_2 are both relatively rare, while in case D, A_1 is common and B_2 is rare (case D is the same as case B, except for the recombination values). With nonzero recombination the $D'_{i1}(i = 2, \ldots, r)$ values decay from their initial value of -1. (Note that for a hitchhiking event to be important in influencing allele frequencies and linkage disequilibrium values we require that the selection intensity exceed the recombination fraction ($s > R$) (THOMSON 1977; ASMUSSEN and CLEGG 1981; ASMUSSEN 1986), and a similar constraint probably obtains for direct selection.)

When one haplotype is favored, as in our model, a total of $r + s - 2$ related haplotypes (haplotypes having an allele in common with the selected haplotype) move into the negative space, while the remainder $(r - 1)(s - 1)$ of haplotypes, that is, all those haplotypes unrelated to the selected haplotype, move into the positive space. If selection is operating on K haplotypes concurrently, and these haplotypes do not share any alleles, we expect $K(r + s) - 2(1 + 2 + \ldots + K)$ haplotypes in the negative linkage disequilibrium space, and the remainder in the positive space. We define the observed number of haplotypes in the negative linkage disequilibrium space by M, and if $b = r + s - 1$, then we obtain the quadratic $K = (b \pm \sqrt{b^2 - 4M})/2$. The distribution of linkage disequilibrium values in the positive and negative spaces can

TABLE 1

Allele frequencies and normalized linkage disequilibria

		Case A		Case B		Case C		Case D	
Generation		0	50	0	50	0	50	0	50
				Allele frequencies					
Alleles	A_1	0.5	0.75	0.5	0.55	0.05	0.12	0.5	0.54
	A_2	0.1	0.05	0.1	0.09	0.1	0.09	0.1	0.09
	A_3	0.15	0.07	0.15	0.13	0.15	0.14	0.15	0.14
	A_4	0.2	0.10	0.2	0.18	0.2	0.19	0.2	0.18
	A_5	0.05	0.02	0.05	0.05	0.5	0.46	0.05	0.05
	B_1	0.05	0.53	0.001	0.10	0.001	0.07	0.001	0.08
	B_2	0.15	0.07	0.049	0.04	0.049	0.05	0.049	0.04
	B_3	0.3	0.15	0.15	0.14	0.15	0.14	0.15	0.14
	B_4	0.5	0.25	0.3	0.27	0.3	0.28	0.3	0.28
	B_5			0.5	0.45	0.5	0.46	0.5	0.46
				Normalized disequilibria (D')					
Haplotypes	1,1	0	0.905	1	1	1	0.885	1	0.888
	1,2	0	−0.334	−0.02	−0.105	−0.388	−0.608	−0.02	−0.078
	1,3	0	−0.334	0	−0.087	0	−0.507	0	−0.067
	1,4	0	−0.334	0	−0.087	0	−0.507	0	−0.067
	1,5			0	−0.087	0	−0.507	0	−0.067
	2,1	0	−0.905	−1	−1	−1	−0.885	−1	−0.888
	3,1	0	−0.905	−1	−1	−1	−0.885	−1	−0.888
	4,1	0	−0.905	−1	−1	−1	−0.885	−1	−0.888
	5,1	0	−0.905	−1	−1	−1	−0.885	−1	−0.888
	2,2	0	0.081	0.002	0.013	0.002	0.008	0.002	0.009
	2,3	0	0.177	0	0.016	0	0.011	0	0.013
	2,4	0	0.334	0	0.039	0	0.026	0	0.030
	2,5			0	0.087	0	0.058	0	0.067
	3,2	0	0.081	0.004	0.02	0.004	0.013	0.004	0.015
	3,3	0	0.177	0	0.016	0	0.011	0	0.013
	3,4	0	0.334	0	0.039	0	0.026	0	0.030
	3,5			0	0.087	0	0.058	0	0.067
	4.2	0	0.111	0.005	0.028	0.005	0.018	0.005	0.021
	4,3	0	0.177	0	0.023	0	0.015	0	0.018
	4,4	0	0.334	0	0.039	0	0.026	0	0.030
	4,5			0	0.087	0	0.058	0	0.067
	5,2			0.001	0.006	0.02	0.070	0.001	0.004
	5,3			0	0.016	0	0.058	0	0.013
	5,4			0	0.039	0	0.058	0	0.030
	5,5			0	0.087	0	0.058	0	0.067

Allele frequencies and normalized linkage disequilibria (D') are presented at generations 0 and 50 with selection coefficient 0.1 (dominant directional selection) for the A_1B_1 haplotype under four contrasting conditions: Case A, all initial disequilibrium values zero and no recombination ($R = 0$); Case B, B_1 is a new mutant on haplotype A_1B_1 which arose from haplotype A_1B_2 with A_1 common and B_2 rare, $R = 0$; Case C, B_1 is a new mutant on haplotype A_1B_1 which arose from haplotype A_1B_2 and A_1 and B_2 are relatively rare, $R = 0.01$; Case D, as for Case B with $R = 0.01$. Related haplotypes are bracketed.

thus give an indication of the minimum number of selection events which are acting, or have acted, on the system.

Neutrality: When mutation creates a new allele at a locus in a population, a new and unique haplotype is formed. This new haplotype will have the maximum positive D' value, 1.0, all related haplotypes with the new mutant allele will be absent and so have D' values of -1.0, while the linkage disequilibrium of the other related haplotypes with the non-mutant allele will be unaltered from their original value, except for the haplotype from which the new mutant derives (see equations (I-15a)–(I-15e) in APPENDIX). If the new haplotype persists in the population and increases in frequency, recombination will generate new related haplotypes in numbers proportional to the frequencies of the alleles at the unmutated locus. It is possible to imagine that genetic drift and mutation alone can produce the same pattern in the disequilibrium space as the selection models described above. Initially, the related haplotypes which have the new mutant allele have disequilibrium values which are maximized and proportional to the frequency of the unshared allele, and identical normalized disequilibrium values, equaling -1. Is the pattern generated by a selection event mimicked as recombination and genetic drift operate to reduce the disequilibrium values from their maximum?

The simulation technique of HUDSON (1983) has been employed to examine the disequilibrium pat-

terns generated in a two locus neutrality model. We generated data samples of size 10,000 for $4Nc$ values of 0, 1, 10, 25, 50, 75, 100, 160 and 500, where N is the effective population size and c is the amount of recombination between loci. (The rarer alleles were combined into one allele class to mimic the undefined "blank" alleles found in HLA data.) We chose $\theta = 4N\mu$, where μ is the mutation rate, so as to generate samples with large numbers of alleles, close to the 13 alleles at one locus and 21 at the other as seen for the HLA data analyzed in the companion paper KLITZ and THOMSON (1987).) Once the normalized disequilibrium values move away from the $D' = -1$ region the pattern generated by a selection event is not mimicked. A very wide range of normalized linkage disequilibrium values was observed and the approximate constancy of normalized linkage disequilibrium values of each set of related haplotypes in the negative disequilibrium space, as predicted by the selection model, is not found in the neutrality case once recombination starts to reduce the initial $D' = -1$ values.

At this stage we have no objective method for determining if a population sample fits the selection pattern or not. The patterns we have described as indicative of selection were arrived at by considering a completely deterministic model. Sampling variance will always cause some deviations from the deterministic predictions. At this exploratory phase of data analysis we have erred on the conservative side and make no statement about data sets which are neither very close to nor very far from the selection predictions. We are currently investigating various approaches to statistical testing.

Migration or continuous admixture: If sizeable differences in allele frequencies exist between two populations, then large amounts of linkage disequilibrium can be generated by migration or admixture (see for example, NEI and LI 1973, FELDMAN and CHRISTIANSEN 1975, THOMSON, BODMER and BODMER 1976). However, except in a very limited set of cases, the pattern of disequilibrium that will be generated by migration or admixture is very different from the specific pattern predicted in the case of selection. This is demonstrated theoretically by considering the standard static model of migration where two populations are mixed in proportions m and $1 - m$, and the case of continuous admixture of the two populations, and then by computer iteration of general migration models.

Consider a two locus, multiallelic model and let the frequency of the allele A_i at the first locus be p_i in population 1, and P_i in population 2, for $i = 1, \ldots, r$. Similarly, let the frequency of the allele B_j at the second locus be q_j in population 1, and Q_j in population 2, for $j = 1, \ldots, s$. In the static mixed population the linkage disequilibrium D_{ij} between the alleles A_i and B_j is given by

$$D_{ij} = mD^1_{ij} + (1 - m)D^2_{ij} + m(1 - m)(p_i - P_i)(q_j - Q_j) \qquad i = 1, \ldots, r; \quad j = 1, \ldots, s, \tag{5a}$$

where D^1_{ij} and D^2_{ij} denote the linkage disequilibria in the two populations considered separately (see CAVALLI-SFORZA and BODMER 1971, page 69). If there is no linkage disequilibrium in each population initially, that is $D^1_{ij} = D^2_{ij} = 0$, then

$$D_{ij} = m(1 - m)(p_i - P_i)(q_j - Q_j) \qquad i = 1, \ldots, r; \quad j = 1, \ldots, s, \tag{5b}$$

so that, $D_{ij} \neq 0$ even if $D^1_{ij} = D^2_{ij} = 0$, provided the allele frequencies in the two populations are different.

Take the case of admixture of say population 2 into population 1 (THOMSON, BODMER and BODMER 1976), where a proportion α of matings each generation involve an individual from population 1 with an individual from population 2, with the progeny entering population 1. If the disequilibria in the two populations are initially zero, then, at generation n the disequilibria in the admixed population are given by

$$D_{ij}(n) = A(p_i - P_i)(q_j - Q_j) \qquad i = 1, \ldots, r; \quad j = 1, \ldots, s \tag{6a}$$

where

$$A = \frac{\alpha}{2}\left[1 - \frac{\alpha}{2} - \frac{R}{2}\right]\left[1 - \frac{\alpha}{2}\right]^{2(n-1)} \left[1 + \sum_{i=1}^{n-1} \frac{(1 - R)^i}{\left(1 - \frac{\alpha}{2}\right)^i}\right] \tag{6b}$$

and R is the recombination fraction between the A and B loci.

The only case of migration or admixture which can mimic the selection pattern with say the haplotype A_1B_1 having high positive linkage disequilibrium and all related haplotypes at both loci having negative disequilibria proportional to the frequency of the unshared allele is the very restrictive situation where alleles A_1 and B_1 are both more frequent in population 1 than population 2, or vice versa, and all other alleles are rarer in population 1 than population 2, or vice versa. We assume that there is no linkage disequilibrium between any pairs of alleles in each population initially. In the static mixed population, and in the continuously admixed population for this example, alleles A_1 and B_1 will be in positive linkage disequilibrium, while all combinations of A_1 with non B_1 alleles, and B_1 with non A_1 alleles, will be in the negative linkage disequilibrium space. The frequencies of the B_j alleles in the mixed (or admixed) population are given by

$$q^*_j = m_n q_j + (1 - m_n)Q_j \qquad j = 2, \ldots, s, \tag{7}$$

where $m_n = m$ in the static migration model and $m_n = [1 - (\alpha/2)]^n$ in the admixture model. The haplotypes in the negative disequilibrium space, say for example A_1 with all non-B_1 alleles, will only have linkage disequilibrium values proportional to the B_j, $j = 2, \ldots, s$, frequency in the mixed population (the selection expectation) if

$$-B(p_1 - P_1)(q_j - Q_j) = \beta[m_n q_j + (1 - m_n)Q_j], \qquad j = 2, \ldots, s,$$

where $B = m(1 - m)$ in the static migration model, and $B = A$ in the admixture model and $\beta > 0$, that is if

$$-(q_j - Q_j) = \gamma[m_n q_j + (1 - m_n)Q_j] \qquad j = 2, \ldots, s,$$

where $\gamma = \beta/B(p_1 - P_1) > 0$, that is, only under the very stringent condition that for all $j = 2, \ldots, s$

$$Q_j = \frac{1 + m_n\gamma}{1 - \gamma(1 - m_n)} q_j. \tag{8}$$

Thus, the situation where the selection pattern will be mimicked by migration occurs only under the most unlikely situation of a strict proportionality of the non-A_1, non-B_1 alleles in the two populations, and also that both A_1 and B_1 are more frequent in one population than the other, and that all the other alleles in that population are all rarer. A simple case of this type would be if the haplotype A_1B_1 is fixed in one of the populations. We have shown by computer iteration that this same exact result applies for continual migration occuring each generation between two populations. In general, the patterns of linkage disequilibrium generated by migration will differ greatly from the selection expectations.

DISCUSSION AND SUMMARY

Disequilibrium pattern analysis is a general method useful for revealing the evolutionary dynamics of tightly linked highly polymorphic loci in natural populations. Recent selection events can be identified from the pattern of the array of two-locus haplotypes in the disequilibrium space subdivided on the basis of all haplotypes sharing one allele. The pattern of haplotypes generated by a selection event is distinct from those produced by migration or random genetic drift. The analysis assumes zero linkage disequilibrium initially for all haplotypes, before the onset of the selection event. The distinct pattern of linkage disequilibrium values generated by a selection event holds for all generations while the selection is acting. This pattern will also presumably be maintained for many generations after the selection event, until the disequilibrium pattern is broken down by recombination and genetic drift. The same basic results continue to apply if the selection involves a new mutant, if we assume zero linkage disequilibrium before the new mutant arose.

In particular, the following criteria are used to reveal selection, and specifically to identify those two-locus haplotypes showing the effect of selection, and thereby differentiated from other haplotypes in the population: (i) the magnitude of the expected frequency and D values of those haplotypes having positive disequilibrium values, (ii) the presence of just one or a few haplotypes in the positive disequilibrium space when we plot the linkage disequilibrium for all haplotypes containing a given allele, and (iii) related haplotypes sharing an allele with a selected haplotype assume a value of linkage disequilibrium, D, proportional to the frequency of the unshared allele and have a common negative value of D'.

In addition, the distribution of the number of haplotypes having positive and negative disequilibrium values can be used to predict the minimum number of recent haplotypic or hitchhiking selection events occurring in a region. The less common of the two alleles in a selected two locus haplotype will have arisen more recently. This conclusion is substantiated by the presence of fewer haplotypes with positive D values in the graph of the linkage disequilibrium values for all the haplotypes with the rarer allele. The disequilibrium pattern based on the rare allele will be the most sensitive in revealing the selective event, as the pattern based on the more common allele will also carry evidence of its previous history.

Disequilibrium pattern analysis is particularly pertinent to the analysis of data from multigene families, marker trait associations and restriction fragment length polymorphism data. The close linkage of a number of loci in the HLA region, and the high level of polymorphism exhibited by many of these loci make it ideally suited for the study of evolutionary forces by disequilibrium pattern analysis. The companion paper (KLITZ and THOMSON 1987) analyzes a large sample of two-locus HLA data in this fashion.

We thank R. HUDSON for the use of his elegant simulation program and P. HEDRICK for his helpful comments on the manuscript. This research was supported by National Institutes of Health grants HD12731 and GM35326 and a UC Berkeley IBM Academic Information Systems grant.

LITERATURE CITED

ALBERT, E., M. BAUR and W. MAYR (Editors), 1984 *Histocompatibility Testing 1984*. Springer-Verlag, Berlin.

ASMUSSEN, M. A., 1986 The dynamics of interlocus associations in the three locus hitchhiking model. 2. The pairwise linkage disequilibrium between two neutral loci. J. Math. Biol. **23:** 285–304.

ASMUSSEN, M. A. and M. T. CLEGG, 1981 Dynamics of the linkage disequilibrium function under models of gene-frequency hitchhiking. Genetics **99:** 337–356.

BAUR, M. P. and D. GRANGE, 1983 Negative linkage disequilibria. Tissue Antigens **21:** 262–263.

CAVALLI-SFORZA, L. L. and W. F. BODMER, 1971 *The Genetics of Human Populations.* W. H. Freeman, San Francisco.

DAUSSET, J. and P. J. COLOMBANI (Editors), 1973 *Histocompatibility Testing 1972.* Munksgaard, Copenhagen.

FELDMAN, M. and F. B. CHRISTIANSEN, 1975 The effect of population subdivision on two loci without selection. Genet. Res. **24:** 151–162.

HANSEN, H. E., S. O. LARSEN, L. P. RYDER and L. S. NIELSON, 1979 HLA-A,B haplotype frequencies in 5,202 unrelated Danes by a maximum-likelihood method of gene counting. Tissue Antigens **13:** 143–153.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

KLITZ, W. and G. THOMSON, 1987 Disequilibrium pattern analysis. II. Application to Danish HLA-A and B locus data. Genetics **116:** 633–643.

LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

NEI, M. and W-H. LI, 1973 Linkage disequilibrium in subdivided populations. Genetics **75:** 213–219.

TERASAKI, P. (Editor), 1980 *Histocompatibility Testing 1980.* University of California, Los Angeles.

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

THOMSON, G., W. F. BODMER and J. BODMER, 1976 The HL-A system as a model for studying the interaction between selection, migration and linkage. pp. 456–498. In: *Population Genetics and Ecology,* Edited by S. KARLIN and E. NEVO. Academic Press, New York.

Communicating editor: W. J. EWENS

APPENDIX

For generality and realism we consider a hitchhiking model, where selection is considered not to be acting directly on the alleles A_i ($i = 1, \ldots, r$) and B_j ($j = 1, \ldots, s$) of the A and B loci, but on a third linked locus where a new selected mutant F has arisen on an A_1B_1 bearing haplotype. We denote the frequency of the A_1B_1 haplotype which carries F by y_{11}, and the frequencies of all the other haplotypes A_iB_j ($i = 1, \ldots, r; j = 1, \ldots, s$), which carry the alternate allele f by x_{ij}, with $y_{11} + \sum_{i=1}^{r} \sum_{j=1}^{s} x_{ij} = 1$. We denote the allele frequency of A_1 by p_{A_1}, of f by p_f, etc. At this stage we assume zero recombination. (Note that the results we obtain below also apply if selection is acting directly on the haplotype A_1B_1).

The selection acting is assumed to be heterozygote advantage (although the results obtained below apply equally to directional or frequency dependent selection), with the following fitness values,

$$\begin{matrix} FF & Ff & ff \\ 1 - s & 1 & 1 - t \end{matrix} \tag{I-1}$$

If haplotype frequencies in the next generation are denoted by primes (′), then

$$\begin{aligned} \bar{w}y'_{11} &= y_{11}^2(1 - s) + y_{11}(1 - y_{11}) \\ &= y_{11}(1 - sy_{11}), \end{aligned} \tag{I-2a}$$

$$\begin{aligned} \bar{w}x'_{ij} &= y_{11}x_{ij} + (1 - t)[x_{ij}^2 + x_{ij}(1 - x_{ij} - \\ &= x_{ij}[1 - t(1 - y_{11})] \end{aligned} \tag{I-2b}$$

($i = 1, \ldots, r;\ j = 1, \ldots, s$), where

$$\bar{w} = 1 - sy_{11}^2 - t(1 - y_{11})^2. \tag{I-2c}$$

From (I-2b), we can write

$$\frac{x'_{ij}}{x_{ij}} = \frac{1 - t(1 - y_{11})}{\bar{w}},$$

and using (I-2a) and (I-2c), this can be written as

$$\frac{x'_{ij}}{x_{ij}} = \frac{1 - y'_{11}}{1 - y_{11}}. \tag{I-3}$$

If haplotype frequencies at generation n are denoted by $y_{11}(n)$ and $x_{ij}(n)$, then it follows from (I-3) that

$$\frac{x_{ij}(n)}{x_{ij}(0)} = \frac{1 - y_{11}(n)}{1 - y_{11}(0)} = \frac{p_f(n)}{p_f(0)}, \tag{I-4}$$

($i = 1, \ldots, r; j = 1, \ldots, s$). Also, from (I-4),

$$\frac{p_{A_i}(n)}{p_{A_i}(0)} = \frac{p_{B_j}(n)}{p_{B_j}(0)} = \frac{p_f(n)}{p_f(0)}. \tag{I-5}$$

($i = 2, \ldots, r; j = 2, \ldots, s$).

We define the linkage disequilibrium between alleles A_i and B_j at generation n by $D_{ij}(n)$, and from equation (1),

$$D_{ij}(n) = x_{ij}(n) - p_{A_i}(n)p_{B_j}(n) \tag{I-6}$$

($i = 1, \ldots, r; j = 1, \ldots, s$; except $i = j = 1$). We denote the change in allele frequencies by

$$\delta_{A_i}(n) = p_{A_i}(n) - p_{A_i}(0) \tag{I-7a}$$

and

$$\delta_{B_j}(n) = p_{B_j}(n) - p_{B_j}(0). \tag{I-7b}$$

Using (I-4) and (I-5) in (I-6) we obtain

$$\begin{aligned} D_{1j}(n) &= x_{1j}(n) - p_{A_1}(n)p_{B_j}(n) \\ &= \frac{p_f(n)}{p_f(0)}[x_{1j}(0) - p_{A_1}(n)p_{B_j}(0)] \\ &= \frac{p_f(n)}{p_f(0)}[p_{A_1}(0)p_{B_j}(0) + D_{1j}(0) - p_{A_1}(n)p_{B_j}(0)] \\ &= -p_{B_j}(n)\delta_{A_1}(n) + \frac{p_f(n)}{p_f(0)} D_{1j}(0) \end{aligned} \tag{I-8a}$$

($j = 2, \ldots, s$).

Similarly

$$D_{i1}(n) = -p_{A_i}(n)\delta_{B_1}(n) + \frac{p_f(n)}{p_f(0)} D_{i1}(0), \tag{I-8b}$$

($i = 2, \ldots, r$).

Using equation (2), and (I-8a) and (I-8b) gives,

$$D_{11}(n) = (1 - p_{B_1}(n))\delta_{A_1}(n) + \frac{p_f(n)}{p_f(0)} D_{11}(0) \tag{I-9a}$$

$$= (1 - p_{A_1}(n))\delta_{B_1}(n) + \frac{p_f(n)}{p_f(0)} D_{11}(0), \tag{I-9b}$$

and

$$D_{ij}(n) = -p_{B_j}(n)\delta_{A_i}(n) + \frac{p_f(n)}{p_f(0)} D_{ij}(0) \tag{I-9c}$$

$$= -p_{A_i}(n)\delta_{B_j}(n) + \frac{p_f(n)}{p_f(0)} D_{ij}(0), \tag{I-9d}$$

($i = 2, \ldots, r; j = 2, \ldots, s$).

If the disequilibrium between A_1 and B_1 at the start of the selection event is zero, or positive, then $D_{11}(n)$ will always be positive, similarly for all the $D_{ij}(n)$ ($i = 2, \ldots, r; j = 2, \ldots, s$) terms, since $\delta_{A_1}(n)$ and $\delta_{B_1}(n)$ are positive, and $\delta_{A_i}(n)$ ($i = 2, \ldots, r$) and $\delta_{B_j}(n)$ ($j = 2, \ldots, s$) are negative. The disequilibrium terms for the haplotypes containing only one of the alleles, A_1 or B_1, of the related haplotypes, that is $D_{1j}(n)(j = 2, \ldots, s)$, and $D_{i1}(n)(i = 2, \ldots, r)$ will all similarly be negative.

Case (a). All initial disequilibrium values are zero: The nega-

tive disequilibria of the related haplotypes A_1 with the B_j's ($j \neq 1$) are all proportional to the present B_j frequency, if the initial disequilibrium values are zero, that is, $D_{1j}(0) = 0$, for $j = 2, \ldots, s$ (from I-8a),

$$D_{1j}(n) = -p_{B_j}(n)\delta_{A_1}(n). \tag{I-10a}$$

Similarly, the negative disequilibria of the related haplotypes B_1 with the A_i's ($i \neq 1$) are all proportional to the present A_i frequency, if the initial disequilibrium values are zero, that is, $D_{i1}(0) = 0$, for $i = 2, \ldots, r$ (from I-8b),

$$D_{i1}(n) = -p_{A_i}(n)\delta_{B_1}(n). \tag{I-10b}$$

For polymorphic loci where $p_{A_1} + p_{B_j} \leq 1$ ($j = 2, \ldots, s$) (BAUR and GRANGE 1983) as is usually the case, for the negative D_{1j} values $D_{1j}(\max) = p_{A_1}p_{B_j}$ (see equations (4d)–(4e)), and similarly for the negative D_{i1} value, $D_{i1}(\max) = p_{A_i}p_{B_1}$. Thus if the initial disequilibrium values are zero, the normalized disequilibrium values $D'_{1j}(n)$ and $D'_{i1}(n)$ are

$$D'_{1j}(n) = \frac{-\delta_{A_1}(n)}{p_{A_1}(n)}, \tag{I-11a}$$

($j = 2, \ldots, s$), and

$$D'_{i1}(n) = \frac{-\delta_{B_1}(n)}{p_{B_1}(n)}, \tag{I-11b}$$

($i = 2, \ldots, r$). The expected $D'_{1j}(n)$ values are equal for all $j = 2, \ldots, s$, and similarly the $D'_{i1}(n)$ values each have the same expectation for $i = 2, \ldots, r$. (These results for $D'_{1j}(n)$ and $D'_{i1}(n)$ do not hold if $p_{A_1} + p_{B_j} > 1$ or $p_{A_i} + p_{B_1} > 1$, respectively.)

If $D_{11}(0) = 0$ and $p_{A_1} < p_{B_1}$, then

$$D'_{11}(n) = \frac{\delta_{A_1}(n)}{p_{A_1}(n)}, \tag{I-12a}$$

while if $p_{A_1} > p_{B_1}$, then

$$D'_{11}(n) = \frac{\delta_{B_1}(n)}{p_{B_1}(n)}. \tag{I-12b}$$

The positive disequilibria of the unrelated haplotypes A_i and B_j ($i = 2, \ldots, r$, $j = 2, \ldots, s$) are proportional to the A_i and B_j frequencies, if the initial disequilibria values are zero (from (I-9c) and (I-9d)),

$$D_{ij}(n) = -p_{B_j}(n)\delta_{A_i}(n) \tag{I-13a}$$

$$= -p_{A_i}(n)\delta_{B_j}(n) \tag{I-13b}$$

If $p_{A_i} < p_{B_j}$, then

$$D'_{ij}(n) = \frac{-\delta_{B_j}(n)}{1 - p_{B_j}(n)}, \tag{I-14a}$$

and if $p_{A_i} > p_{B_j}$, then

$$D'_{ij}(n) = \frac{-\delta_{A_i}(n)}{1 - p_{A_i}(n)}. \tag{I-14b}$$

The normalized disequilibrium values for haplotypes with a particular allele A_i with all alleles B_j, $j = 2, \ldots, s$, will not usually be equal, depending on the relative allele frequencies. Note that for highly polymorphic systems such as HLA, with a fairly even distribution of allele frequencies, $D'_{ij}(n)$ will be considerably less in value than $D'_{i1}(n)$, since $\delta_{A_1}(n)$ will be larger than $-\delta_{B_j}(n)$, and $p_{A_1}(n)$ will be smaller than $1 - p_{B_j}(n)$ (see (I-12a) and (I-14a) for example).

Case (b). The case where B_1 is a new mutant: We now consider the case where we suppose that B_1 is a newly arisen mutant which forms the haplotype A_1B_1 from the haplotype A_1B_2, and that the haplotype A_1B_1 is selected. The occurrence of the new mutant creates linkage disequilibrium. For comparison with the results above, we assume that before the new mutant arises all pairwise disequilibrium values are zero. The linkage disequilibrium values created by the occurrence of the new mutant are as follows:

$$D_{11}(0) = p_{B_1}(0)(1 - p_{A_1}(0)) \tag{I-15a}$$

$$D'_{11}(0) = 1$$

$$D_{i1}(0) = -p_{A_1}(0)p_{B_1}(0) \tag{I-15b}$$

$$D'_{i1}(0) = -1 \quad i = 2, \ldots, r$$

$$D_{12}(0) = -p_{B_1}(0)(1 - p_{A_1}(0)) \tag{I-15c}$$

$$D'_{12}(0) = -p_{B_1}(0)(1 - p_{A_1}(0))/[p_{A_1}(0)p_{B_2}(0)]$$

$$D_{i2}(0) = p_{A_i}(0)p_{B_1}(0) \tag{I-15d}$$

$$D'_{i2}(0) = \begin{cases} p_{B_1}(0)/[1 - p_{B_2}(0)] & \text{if } p_{A_i}(0) < p_{B_2}(0) \\ p_{A_i}(0)p_{B_1}(0)/[p_{B_2}(0)(1 - p_{A_1}(0))] & \text{if } p_{A_i}(0) > p_{B_2}(0) \quad i = 2, \ldots, r \end{cases} \tag{I-15e}$$

$$D_{1j}(0) = 0$$

$$D'_{1j}(0) = 0 \quad j = 3, \ldots, s$$

and all other linkage disequilibrium values zero (hence all the normalized values).

The $D_{i1}(n)$, $i = 2, \ldots, r$, values still stay proportional to the current A_i allele frequencies, with

$$D_{i1}(n) = -p_{A_i}(n)p_{B_1}(n), \tag{I-16}$$

and this should be compared with (I-10b). In this case $D'_{i1}(n) = -1$, as recombination is assumed zero. Iterations of deterministic selection models with recombination, show that the linkage disequilibrium and normalized values with recombination yield the same general results obtained above for the case of no recombination. That is, the $D_{i1}(n)$ values are proportional to the A_i allele frequencies and the $D'_{i1}(n)$ values are all equal.

For the A_1B_j haplotypes, excluding A_1B_2, that is $j = 3, \ldots, s$, the initial linkage disequilibrium values are all zero (hence also the normalized values) and all the results obtained above for the proportionality of the disequilibrium values and equality of the normalized values for the case of zero initial disequilibrium continue to apply. The haplotype A_1B_2 from which the mutant arose, will show a different pattern, as the initial disequilibrium (and normalized disequilibrium) are nonzero. If the frequencies of A_1 and B_2 are reasonably large relative to the frequency of the new mutant B_1 (so that $D'_{12}(0)$ is small) the $D_{12}(n)$ and $D'_{12}(n)$ values will not deviate greatly from the case when all the initial disequilibrium values are zero. However, if the new mutant arises on a relatively rare haplotype, then $D'_{12}(0)$ will be a large negative value, and the values of $D_{12}(n)$ and $D'_{12}(n)$ may deviate considerably from the patterns of $D_{1j}(n)$ and $D'_{1j}(n)$, especially in the initial generations of selection.

Case (c). The effects of recombination: The general equations to determine haplotype frequencies in the next generation (x'_{ij}), given haplotype frequencies in this generation (x_{ij}), fitness parameters w_{ijkl} for individuals of genotype $x_{ij}x_{kl}$, ($i, k = 1, \ldots, r$; $j, l = 1, \ldots, s$) and recombination R between the A and B loci are

$$\bar{w}x'_{ij} = \sum_{k=1}^{r}\sum_{l=1}^{s} w_{ijkl}x_{ij}x_{kl} - R\left[\sum_{\substack{k=1\\k\neq i}}^{r}\sum_{\substack{l=1\\l\neq j}}^{s} w_{ijkl}x_{ij}x_{kl} - \sum_{\substack{k=1\\k\neq i}}^{r}\sum_{\substack{l=1\\l\neq j}}^{s} w_{ilkj}x_{il}x_{kj}\right] \tag{I-17a}$$

where

$$\bar{w} = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{r}\sum_{l=1}^{s} w_{ijkl}x_{ij}x_{kl}. \tag{I-17b}$$

If we consider dominant directional selection on the haplotype A_1B_1 (fitness $1 + s$ compared to fitness of 1 for other haplotypes), then

$$\bar{w}x'_{11} = x_{11}(1 + s) - R\,[(1 + s\chi_{11})\,D_{11} + s\chi_{11}\,(1 - p_{A_1})\,(1 - p_{B_1})] \tag{I-18a}$$

$$\bar{w}x'_{1j} = x_{1j}[1 + s(1 - R)x_{11}] - R(D_{1j} - sx_{11}p_{B_j}), \quad j = 2, \ldots, s, \tag{I-18b}$$

$$\bar{w}x'_{i1} = x_{i1}[1 + s(1 - R)x_{11}] - R(D_{i1} - sx_{11}p_{A_i}), \quad i = 2, \ldots, r, \tag{I-18c}$$

$$\bar{w}x'_{ij} = x_{ij}[1 + s(1 - R)x_{11}] - RD_{ij}, \quad i = 2, \ldots, r; \quad j = 2, \ldots, s, \tag{I-18d}$$

where

$$\bar{w} = 1 + 2sx_{11} - sx_{11}^2. \tag{I-18e}$$

Iteration of deterministic selection models with recombination shows that all the results obtained above still apply.