Selection, Hitchhiking and Disequilibrium Analysis at Three Linked Loci With Application to HLA Data

Wendy P. Robinson,*,[2] Anne Cambon-Thomsen,† Nicolas Borot,† William Klitz* and Glenys Thomson*,[1]

**Department of Integrative Biology, University of California, Berkeley, California 94720, and †INSERM U100, C.H.U. Purpan, 31052 Toulouse Cedex, France*

Manuscript received March 14, 1991
Accepted for publication July 24, 1991

ABSTRACT

The HLA system has been extensively studied from an evolutionary perspective. Although it is clear that selection has acted on the genes in the HLA complex, the nature of this selection has yet to be fully clarified. A study of constrained disequilibrium values is presented that is applicable to HLA and other less polymorphic systems with three or more linked loci, with the purpose of identifying selection events. The method uses the fact that three locus systems impose additional constraints on the range of possible disequilibrium values for any pair of loci. We have thus examined the behavior of the normalized pairwise disequilibrium measures using two locus (D'), and also three locus (D''), constraints on pairwise disequilibria in a three locus system when one of the three loci is under positive selection. The difference between these measures, $\delta = |D'| - |D''|$, has a distribution for the two unselected loci differing from that for the selected locus with either of the unselected loci (the hallmark is a high positive value of δ for the two unselected loci). An examination of genetic drift indicates that positive δ values are unlikely to be found in human populations in the absence of selection when recombination is greater than about 0.1%. This measure can thus provide insight into which allele of several linked loci might have been subject to selection. Application of this method to HLA haplotypes from a large French population study (Provinces Francaise) identifies selected alleles on particular haplotypes. Application of a complementary method, disequilibrium pattern analysis also confirms the action of selection on these haplotypes.

DISEQUILIBRIUM between linked loci may be created by a variety of population-genetic processes. A new mutant begins in maximum positive disequilibrium with the haplotype it arises on, and is in maximum negative disequilibrium with all other haplotypes. Recombination operates to reduce disequilibrium. Genetic drift, subdivided populations, admixture, assortative mating, and selection can also create or maintain disequilibria. Although it has been suggested that the presence of disequilibrium could be a useful indicator of past selection, it is often difficult to distinguish disequilibria due to selection acting directly on loci versus disequilibrium between neutral loci created by a hitchhiking event (THOMSON 1977).

HUDSON (1985) used Monte-Carlo simulations to examine the distributions of several different measures of two-locus disequilibrium under an infinite-allele model with constant population size. The result indicates that in a two-locus, diallelic system with $4Nc < 10$ (N is the effective population size and c is the recombination fraction), it is very unlikely that one would reject a neutral model. It is estimated that there is on average less than one gene per 20–30 kb of DNA in humans (MCKUSICK 1988) (approximately $c = 0.0002$–0.0003). Assuming an effective human population size of 5000–10,000 (NEI and GRAUR 1984), this means that two-allele pairwise disequilibrium analysis may not be useful over a distance much less than that which typically separates two genes. Even for higher values of $4Nc$, the variance of the expected disequilibrium value under a neutral model is quite large. For this reason, it is desirable to examine the properties of systems with either multiple alleles or multiple loci, where more information can be extracted. Applying statistical methods in such systems may, however, be difficult.

HEDRICK and THOMSON (1986) determined the distributional properties of two-locus associations with multiple alleles at each locus using the program of HUDSON (1985). They determined the expected disequilibrium under neutrality for a given sample size and number of alleles in the sample. The variance of the expected disequilibrium value under neutrality tends to decrease as the number of alleles increases. This method applied to South American Indian histocompatibility locus (HLA) data indicated that HLA-

[1] To whom reprint requests should be sent.
[2] Present address: Institute for Medical Genetics, University of Zürich, Raemistrasse 74 CH8001 Zürich, Switzerland.

A and HLA-B locus associations are greater than expected under neutrality (although not significantly so).

THOMSON and KLITZ (1987) examined the distribution of disequilibrium values for one specified allele at a locus with all alleles at another locus (this method is termed disequilibrium pattern analysis or DPA). The distribution of disequilibrium values when the allele of interest has been under selection differs from the pattern expected under random mutation and drift. This distribution is most informative when the loci are highly polymorphic (as with HLA). Examination of Danish HLA-A and HLA-B locus disequilibria patterns identified six A-B combinations (haplotypes) that show patterns indicating selection has occurred (KLITZ and THOMSON 1987).

A number of other observations also indicate that strong selection is acting on the HLA region, including: the extensive polymorphism (see *e.g.*, ALBERT, BAUR and MAYR 1984, DUPONT 1989) with very even allele frequencies (KLITZ, THOMSON and BAUR 1986); the higher frequency of nonsynonomous amino acid changes than silent changes in the antigen recognition site (ARS) *vs.* the excess of silent changes in the rest of the molecule (HUGHES and NEI 1988, 1989); the preferential occurrence of high levels of variability at positions critical to antigen recognition (HEDRICK, WHITTAM and PARHAM 1991; HEDRICK *et al.* 1991); and the great age of alleles (FIGUEROA, GUNTHER and KLEIN 1988; LAWLOR *et al.* 1988; GYLLENSTEN and ERLICH 1989; KUHNER *et al.* 1990, 1991; GYLLENSTEN, LASHKARI and ERLICH 1990). The frequency distribution of immunologically detectable alleles (KLITZ, THOMSON and BAUR 1986) and the preferential occurrence of nonsynonymous substitutions in the antigen binding regions of the HLA class I and II loci (HUGHES and NEI 1988, 1989) clearly show that selection has specifically acted to maintain functional diversity at these loci. The large number and age of the alleles have, as yet, best fit with models of overdominant selection (TAKAHATA and NEI 1990). Various types of viability selection, maternal-fetal selection, and non-random mating may also influence HLA variation (HEDRICK *et al.* 1991) and it thus seems likely that multiple mechanisms are responsible for HLA diversity. It should also be noted that the allele frequency distribution of loci in the class III or complement region, which lies between the tightly linked HLA class I and II regions, appear to either fit neutrality expectations or purifying selection (KLITZ, THOMSON and BAUR 1986).

As the exact nature of the selection in this region has yet to be fully clarified, and is probably multifaceted, it is useful to examine HLA data in as many complementary ways as possible to detect clues of past and present selection events. We introduce the approach of examining pairwise disequilibrium values in the context of three locus systems as a useful example of exploratory data analysis.

In the present study, patterns of constrained (three-locus) and unconstrained (two-locus) pairwise disequilibria under a hitchhiking-selection model and under genetic drift are examined. This study of disequilibrium patterns is applicable to systems with three, or more, linked loci with two, or more, alleles at each locus. (The high level of polymorphism seen with HLA is not required for this approach, but is required for disequilibrium pattern analysis) (THOMSON and KLITZ 1987). In a companion study (ROBINSON, ASMUSSEN and THOMSON 1991), it was shown that a three-locus system imposes additional constraints, over the constraints that exist when considering two loci alone, on the range of possible disequilibrium values (D) for any pair of loci.

It is often of interest to consider the normalized disequilibrium measure D', which is equal to the pairwise D value divided by the maximum (or minimum if D is negative) value it can take given the allele frequencies at the two loci (LEWONTIN 1964). For a three (or more) locus system, a new normalized constrained pairwise measure, D'', was defined as the observed pairwise D value divided by its range of possible positive values (if $D > 0$) or possible negative values (if $D < 0$) in the three locus constrained system (ROBINSON, ASMUSSEN and THOMSON 1991).

In the present study, the behavior of the pairwise normalized measures D' and D'' for each combination of loci in a three-locus system, when one of the three loci is under positive selection, is examined. The impetus for this study was the observation with the HLA A1-B8-DR3 haplotype that the D' and D'' values for A1-DR3 were quite different in magnitude and direction than the values for the other two pairs of alleles (ROBINSON, ASMUSSEN and THOMSON 1991). This observation led to the hypothesis that although a "hitchhiking" selection event (THOMSON 1977) can create linkage disequilibrium between neutral loci, this disequilibrium may be closer to its minimum possible in the three locus (constrained) system, such that $|D''| < |D'|$ (*i.e.*, $\delta = |D'| - |D''| > 0$, *e.g.*, A1-DR3) reflecting the fact that these alleles are passengers in the putative "hitchhiking" event. In contrast, the disequilibrium generated between a neutral and selected locus may be reflected in the examples of A1-B8 and B8-DR3 with $|D''| > |D'|$ (*i.e.*, $\delta < 0$). Many combinations of selection coefficients and recombination values are considered to determine what range of patterns are expected when one allele is selected. Analysis of patterns expected under admixture, sampling, and drift is also explored for comparison.

This approach of considering constrained pairwise disequilibrium values in a three locus system is then

applied to all three-way combinations of seven loci on the fourteen most common HLA haplotypes from a large French population study consisting of 5460 haplotypes (CAMBON-THOMSEN *et al.* 1986). The method of disequilibrium pattern analysis (THOMSON and KLITZ 1987; KLITZ and THOMSON 1987; KLITZ *et al.* 1991) is also applied to the data set. The constrained disequilibrium values of specific haplotypes indicate that alleles *B7, B8, B13, B35, B57, BfF1, BfSO7* and *DR4* show some signs of past selection. Except for the two rare *Bf* alleles, and weakly for *DR4*, the haplotypes associated with each of these alleles also show evidence of selection using disequilibrium pattern analysis.

METHODS

Disequilibrium measures: The D, D' and D'' measures of pairwise disequilibrium are used. D is the usual measure of pairwise disequilibrium, with $D_{AB} = f(AB) - p_A p_B$ where $f(AB)$ is the frequency of the AB haplotype and p_A and p_B denote allele frequencies at the A and B loci. D' and D'' are normalized pairwise disequilibrium measures: D' is the normal two locus measure of normalized disequilibrium, D'' is the normalized pairwise disequilibrium which allows for the additional constraints placed by a third locus on the possible values of the pairwise disequilibrium.

In a two-locus system, the constraints on the pairwise disequilibrium term are defined by the allele frequencies:

$$-p_A p_B,\ -q_A q_B \leq D_{AB} \leq p_A q_B,\ q_A p_B \tag{1}$$

$$\max D_{AB} = \min(p_A q_B,\ q_A p_B) \tag{2a}$$

$$\min D_{AB} = \max(-p_A p_B,\ -q_A q_B) \tag{2b}$$

$$\min D_{AB} = \max(-p_A p_B,\ -q_A q_B)$$

where ($q_A = 1 - p_A$ and $q_B = 1 - p_B$).

When considering three loci, the additional constraints on D_{AB} are functions of the three allele frequencies and the other two pairwise disequilibria (see ROBINSON, ASMUSSEN and THOMSON 1991 for details):

$$-p_A p_B,\ -q_A q_B,\ -m_1,\ -m_2 \leq D_{AB} \leq p_A q_B,\ q_A p_B,\ M_1,\ M_2 \tag{3}$$

where

$$m_1 = p_A p_B p_C + q_A q_B q_C + D_{AC} + D_{BC} \tag{4a}$$

$$m_2 = p_A p_B q_C + q_A q_B p_C - D_{AC} - D_{BC} \tag{4b}$$

$$M_1 = p_A q_B p_C + q_A p_B q_C + D_{AC} - D_{BC} \tag{4c}$$

$$M_2 = p_A q_B q_C + q_A p_B p_C - D_{AC} + D_{BC}. \tag{4d}$$

$$\min {}^*D_{AB} = \max\{-p_A p_B,\ -q_A q_B,\ -m_1,\ -m_2\} \tag{5a}$$

$$\max {}^*D_{AB} = \min\{p_A q_B,\ q_A p_B,\ M_1,\ M_2\}. \tag{5b}$$

The normalized disequilibrium D' defined by LEWONTIN (1964) is the observed D value divided by its maximum, if positive, or minimum, if negative, possible value in the two locus system, with a range from -1 to $+1$.

$$D'_{AB} = \begin{cases} \dfrac{D_{AB}}{\max D_{AB}} & \text{if } D_{AB} > 0 \\ \dfrac{D_{AB}}{-\min D_{AB}} & \text{if } D_{AB} < 0 \\ 0 & \text{if } D_{AB} = 0. \end{cases} \tag{6}$$

We have defined D'', the normalized disequilibrium value which allows for the constraints imposed by the three-locus system, to be the amount by which the observed D value exceeds the minimum possible of that sign, divided by its range of possible values of that sign (ROBINSON, ASMUSSEN and THOMSON 1991). Since max $*D_{AB}$ can be less than zero, and min $*D_{AB}$ can be greater than zero, we must allow for these possibilities in our definition of D''_{AB}.

$$D''_{AB} = \begin{cases} \dfrac{D_{AB}}{\max {}^*D_{AB}} & \text{if } D_{AB} > 0 \text{ and } \min {}^*D_{AB} \leq 0 \\ \dfrac{D_{AB} - \min {}^*D_{AB}}{\max {}^*D_{AB} - \min {}^*D_{AB}} & \text{if } D_{AB} > 0 \text{ and } \min {}^*D_{AB} > 0 \\ \dfrac{D_{AB}}{-\min {}^*D_{AB}} & \text{if } D_{AB} < 0 \text{ and } \max {}^*D_{AB} \geq 0 \\ \dfrac{D_{AB} - \max {}^*D_{AB}}{\max {}^*D_{AB} - \min {}^*D_{AB}} & \text{if } D_{AB} < 0 \text{ and } \max {}^*D_{AB} < 0 \\ 0 & \text{if } D_{AB} = 0. \end{cases} \tag{7}$$

As a measure of the effects (magnitude and direction) of the three-locus constraints relative to the normal two-locus constraints, δ is defined as

$$\delta = |D'| - |D''|. \tag{8}$$

The values of D' and D'' for a particular pair of loci will always be of the same sign, and thus δ measures the deviation of D'' from D' relative to zero.

Model: The selection model examined is as in THOMSON (1977) and assumes that there are three loci, with two alleles at each locus: labeled A and a, B and b, and C and c, in that order on the chromosome. (Multiallelic systems are accomodated by combining all alleles aside from the one under consideration at that locus.) The starting conditions assume that a new selectively advantageous allele arises at locus B or C (a new allele arising at A is equivalent to that arising at C when the recombination distances $A - B$ and $B - C$ are reversed). Because drift is generally more important than selection when allele frequencies are small, it is also assumed that selection only acts after the new

TABLE 1

Summary of starting conditions for each case examined

Case	q_A	q_B	q_C	D_{AB}	D_{AC}	D_{BC}	D_{ABC}	Selected locus
1	0.05	0.01	0.30	0.0095	0.0	0.007	−0.00665	*B*
2	0.05	0.30	0.01	0.0	0.0095	0.007	−0.00665	*C*
3	0.30	0.01	0.30	0.007	0.0	0.007	−0.0049	*B*
4	0.3	0.3	0.01	0.0	0.007	0.007	−0.0049	*C*
5	0.05	0.01	0.5	0.0095	0.0	0.005	−0.00475	*B*
6	0.05	0.5	0.01	0.0	0.0095	0.005	−0.00475	*C*
7	0.15	0.01	0.15	0.0085	0.0	0.0085	−0.007225	*B*
8	0.15	0.15	0.01	0.0	0.0085	0.0085	−0.007225	*C*
9	0.3	0.05	0.01	0.0	0.007	0.0095	−0.00665	*C*
10	0.05	0.01	0.30	0.0	0.0	0.0	0.0	*B*
11	0.05	0.30	0.01	0.0	0.0	0.0	0.0	*C*
12	0.1	0.01	0.01	0.0	0.0	0.0	0.0	*B*
13	0.1	0.01	0.01	0.0	0.0	0.0	0.0	*C*
14	0.01	0.1	0.3	0.0	0.0	0.0	0.0	*B*
15	0.01	0.3	0.1	0.0	0.0	0.0	0.0	*C*
16	0.05	0.3	0.01	0.0	0.0095	0.007	−0.00665	*B*
17	0.05	0.01	0.3	0.0095	0.0	0.007	−0.00665	*C*
18	0.01	0.05	0.3	0.0095	0.007	0.0	−0.00665	*B*
19	0.3	0.01	0.05	0.0	0.007	0.0095	−0.00665	*C*

mutant has reached a frequency of 0.01. Alleles at the two other neutral loci begin in linkage equilibrium (D, D', $D' = 0$) while the new mutant is in maximum disequilibrium with each of the other two loci (D', $D'' = 1$) [see THOMSON (1977) for further details]. As a consequence of the hitchhiking event linkage disequilibrium between the two neutral loci arises if the recombination fraction between the neutral and selected loci is smaller than the order of magnitude of the selective difference at the selected locus.

The new mutant (located at either the *B* locus or the *C* locus) is represented as the lower case allele (*b* or *c*) with frequency q_B or q_C respectively; and the original haplotype carrying the new mutant is always the *a-b-c* haplotype. The situation modeled is that with two alleles at each locus such that the new mutant arises at a previously monomorphic locus.

The fitness of *BB* (or *CC* when *C* is the selected locus) is $1 - s_1$, the fitness of $Bb = 1$, and the fitness of $bb = 1 - s_2$, where s_1 and $s_2 \geq 0$. Both directional selection on a dominant allele (with *b* dominant over *B*) ($s_2 = 0$) and heterosis ($s_1 \geq s_2 > 0$) are modeled. The recombination distance between *A* and *B* is R_1, the recombination between *B* and *C* is R_2 and we assume no interference. The starting conditions examined are listed in Table 1 and consist of cases of a newly arisen selectively advantageous mutant as well as cases where the disequilibrium values are initially equal to zero and cases where a common allele is selected. The recursion system (a deterministic model) relating the gametic frequencies in the next generation to the present one (no overlapping generations) follows THOMSON (1977).

The changes in allele frequencies, three measures of pairwise disequilibrium (D, D', D'', and hence $\delta = |D'| - |D''|$) and two measures of three-way disequilibrium (D_{ABC}, D'_{ABC}) were examined. D'_{ABC} is a normalized measure of three-locus disequilibrium (see THOMSON and BAUR 1984). Selection coefficients used ranged from 0 to 0.1; the range of recombination values used was 0.00001 to 0.01 (for examples see Tables 2–5). Some exceptions to the condition of a newly selectively advantageous mutant were made to investigate the case of selection not acting immediately after a new mutant arises (*i.e.*, if the environment changes such that a previously neutral allele now has a selective advantage). In cases 10–15 the three loci begin in linkage equilibrium (including the rare allele) and for cases 14–19, the selected allele is not the rare allele.

Sampling error and genetic drift: The effect of sampling error is evaluated by taking 1000 randomly drawn samples of size 100, 500 or 1000 from populations with a given set of allele frequencies and disequilibrium values, and calculating the distribution of D', and δ values which result.

Obtaining estimates of the expected values of D'' or δ under genetic drift, *i.e.*, random fluctuations of haplotype frequencies due to a finite population size, is quite difficult because of the number of variables involved. In order to examine the properties of these values under a few specific models relevant to the human HLA loci, genetic drift was examined for the same starting conditions as used for the selection model (see Table 1). Once the population was initiated with specific allele frequencies, disequilibrium values,

TABLE 2

Examples of patterns of δ_{uus}, δ_{us} and δ_{us}' seen for various starting conditions

				Selected locus							
				B	*C*	*B*	*C*	*B*	*C*	*B*	*C*
R_1	R_2	s_1	s_2	case 1	case 2	case 3	case 4	case 5	case 6	case 7	case 8
0.005	0.005	0.1	0.1	I	II	I	III	I, II	I, II	I	I, III
		0.1	0	I	II	I	III	I, II	I, II	I	II′
		0.01	0.01	I	III	O	O	O	I	I	III
		0.01	0	I	III	I	O	O	I	I	III
		0.001	0.001	O	O	O	O	O	O	O	O
		0.001	0	O	O	O	O	O	O	O	O
0.005	0.001	0.1	0.1	I	I	I, III	I, III	I	I	I, III	I, III
		0.1	0	I	II	I, III	I, III	I	I	I, III	I, II′, III
		0.01	0.01	I	I	III	III	I	I	III	III
		0.01	0	I	I	III	III	I	I	III	III
		0.001	0.001	O	I	O	O	O	O	O	O
		0.001	0	O	O	O	O	O	O	O	O
0.001	0.005	0.1	0.1	I, III	II′	I, III	III	II, I	III	I, III	II′
		0.1	0	I, III	II′	I, III	III	II, I	III	I, III	II′
		0.01	0.01	III	I	III	O	III	O	III	I
		0.01	0	III	III	III	O	III	O	III	I
		0.001	0.001	O	O	O	O	O	O	O	O
		0.001	0	O	O	O	O	O	O	O	O
0.001	0.001	0.1	0.1	I	II	I	II, II	II	II	I	II′
		0.1	0	I	II′	II, I	II, III	II, I	II	I	II′
		0.01	0.01	I	I	III	O	I	O	I	I
		0.01	0	I	I	O	O	O	O	I	I
		0.001	0.001	I	O	O	O	O	O	O	O
		0.001	0	O	O	O	O	O	O	O	O
0.001	0.0001	0.1	0.1	I	I	II, III	II′, III	I	I	I	I
		0.1	0	I	II′, III	II, I	II	I	I	I, III	II′, III
		0.01	0.01	I	I	III	III	I	I	I, III	I, III
		0.01	0	I	I	III	III	I	I	I, III	I, III
		0.001	0.001	I	I	O	O	I	I	O	I
		0.001	0	I	I	O	O	I	I	O	I
0.0001	0.001	0.1	0.1	II, I	II′	II, III	II	II	II′	I	II′
		0.1	0	II, I	II′	II, I	II′	II, I	II′	I, III	II′
		0.01	0.01	I, III	II′	III	O	II, III	O	I, III	I
		0.01	0	I, III	II′	III	O	II, I	O	I, III	II
		0.001	0.001	III	O	O	O	O	O	O	O
		0.001	0	III	O	O	O	O	O	O	O
0.0001	0.0001	0.1	0.1	II, I	II	I	II, III	II	II	I	II′
		0.1	0	I	II′	II, I	II′	I	II	I	II′
		0.01	0.01	II, I	II	I	II, III	II	II	I	II′
		0.01	0	II, I	II′	II, I	II′	II, I	II	I	II′
		0.001	0.001	I	I	O	O	I	O	I	I
		0.001	0	I	I	O	O	I	O	I	I

Pattern I, $\delta_{uu} \geq 0 \geq \delta_{us}$; (also includes patterns where one Δ_{us} is only slightly positive but much smaller than Δ_{uu}, *i.e.*, < 1/10th). Pattern II, $\delta_{uu} > \delta_{us} > 0$; pattern II′ is where the situation starts as $\delta_{uu} > \delta_{us} > 0$ but becomes $\delta_{us} > \delta_{uu} > 0$. Pattern III, δ_{uu}, δ_{us}, $\delta_{us}' \leq 0$. Pattern O: all δ values equal zero (or are very close to zero). Refer to Table 1 for the conditions of each case. R_1 and R_2 are the *A-B* and *B-C* recombination fractions; s_1 and s_2 are the selection coefficients (see text).

recombination values (0.0001–0.01), and population size (N), a random sample of $2N$ haplotypes was taken each generation for 1,000 generations or until one of the six alleles (two alleles from each of the three loci) was lost, whichever came first.

An effective diploid population size of 5,000 was used for the simulations. This is a conservative estimate for the effective human population size which has been estimated as about 10,000 (NEI and GRAUR 1984). It has been pointed out that "an effective population size, if one exists, would not be the same for all pairs of sites" (WEIR and HILL 1986). However, data on the correlation between disequilibrium of HLA loci *vs.* map distance supports an effective population size for the HLA loci to be between 5,000–10,000 (ROBINSON 1989). Recombination values in the range of 0.0001–0.01 were used. For each set of conditions, the simulation was repeated 200–500

TABLE 3

Examples of patterns of δ_{uu}, δ_{us} and δ_{us}' seen for $s_1 = 0.01$ and $s_2 = 0$

		Selected locus							
R_1	R_2	*B* case 1	*C* case 2	*B* case 3	*C* case 4	*B* case 5	*C* case 6	*B* case 7	*C* case 8
0.005	0.005	I	III	I	O	O	I	I	III
	0.002	I	I	III	III	I	I	III	III
	0.001	I	I	III	III	I	I	III	I
0.002	0.005	III	O	III	O	III	O	III	III
	0.002	I	I	I	III	I	I	I	III
	0.001	I	I	III	III	I	I	I	III
0.001	0.005	III	I	III	O	III	O	III	I
	0.002	I	I	III	O	I	I	I	III
	0.001	I	I	O	O	I	O	I	I
	0.0001	I	I	III	III	I	I	I, III	I, III
	0.00001	I, III	I	I, III	III	I	I	I, III	I, III
0.0001	0.005	I	I	III	III	III	I	III	I
	0.001	I, III	II′	III	O	II, I	O	I, III	II
	0.0001	I	II′	II, I	II′	II, I	II	I	II′
	0.00001	I	I, II	II, I	II	I	I	I	I, II
0.00001	0.001	I, III	II′	I, III	O	II, III	II′	III	I, III, II
	0.0001	II, I	II′	II, I	II′	II, I	II′	I	II′
	0.00001	II, I	II	II	II	II	II	I	II′

times. As expected, the rare mutant was often lost from the population quite quickly. The goal of this analysis was to examine a large number of cases where the two alleles at each of the three loci were maintained in the population, and determine what patterns of D' and D'' disequilibrium values are observed, and, more specifically, how often $\delta = |D'| - |D''| > 0$, since this state appears to be indicative of selection.

Admixture: The mixing of populations with different allele frequencies may create linkage disequilibrium. Therefore the resulting δ values in various admixed populations were examined where the two mixed populations differed in allele frequencies at three loci, each of which considered alone is in linkage equilibrium. The two initial populations were mixed in equal proportions.

Disequilibrium pattern analysis: A method termed disequilibrium pattern analysis (THOMSON and KLITZ 1987) has previously proven useful in determining selection events in the HLA region (KLITZ and THOMSON 1987). 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.

Disequilibrium pattern analysis is a general method useful for revealing the evolutionary dynamics of tightly linked highly polymorphic loci. The following criteria are used to reveal selection, and specifically to identify the particular two-locus haplotypes showing the effect of positive selection: (i) magnitude of the expected frequency under linkage equilibrium and D values of those haplotypes which have positive disequilibrium, (ii) the presence of just one or a few haplotypes in the positive disequilibrium space when linkage disequilibrium is plotted versus expected haplotype frequency for all haplotypes containing a given allele and (iii) haplotypes sharing an allele with a selected haplotype assume disequilibrium (D) values proportional to the frequency of the unshared allele and have a common negative D' value.

Data: HLA haplotype information was available from the "Provinces Francaises" study (CAMBON-THOMSEN *et al.* 1986). This data was collected from healthy French families in 16 different regions of France, as well as Corsica and Quebec. Haplotypes were assigned for 1,362 families (5460 haplotypes) using the FAP program (NEUGEBAUER, WILLIAMS and BAUR 1984). When haplotype phase is unclear from family information, this program assigns probabilities to all possible haplotypes by an iterative procedure using information on mode of inheritance of alleles at each locus, transmission pattern in the pedigree, and disequilibrium values computed from the population under analysis. As all possible haplotypes are given a probability, a fraction of haplotypes constructed by FAP had estimated population frequencies of less than one individual. Most of these rare haplotypes probably do not actually exist in the population. All haplotypes with frequencies less than 0.0001 (or approximately 0.5 of an individual and summing to a total frequency of 0.0363) have thus been eliminated, and the remaining haplotype frequencies were adjusted to account for this.

TABLE 4

Maximum δ_{uu} and δ_{us} values (listed in that order) reached under various starting conditions

				Selected locus															
				B case 1		*C* case 2		*B* case 3		*C* case 4		*B* case 5		*C* case 6		*B* case 7		*C* case 8	
R_1	R_2	s_1	s_2																
0.005	0.005	0.1	0.1	0.39	0.00	0.23	0.07	0.00	0.00	0.00	0.00	0.23	0.04	0.09	0.01	0.27	0.00	0.15	0.01
		0.1	0	0.42	0.00	0.25	0.07	0.04	0.00	0.00	0.00	0.27	0.05	0.10	0.02	0.30	0.00	0.17	0.12
		0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.01	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.001	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.001	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.005	0.001	0.1	0.1	0.51	0.03	0.44	0.03	0.05	0.00	0.03	0.00	0.31	0.02	0.25	0.01	0.39	0.01	0.33	0.01
		0.1	0	0.54	0.04	0.46	0.10	0.12	0.00	0.08	0.00	0.35	0.02	0.27	0.02	0.44	0.02	0.37	0.13
		0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.01	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.001	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.001	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.001	0.005	0.1	0.1	0.46	0.00	0.27	0.18	0.05	0.00	0.00	0.00	0.29	0.17	0.12	0.10	0.39	0.01	0.19	0.13
		0.1	0	0.48	0.00	0.28	0.20	0.12	0.00	0.00	0.00	0.31	0.15	0.14	0.11	0.44	0.02	0.21	0.19
		0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.01	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.001	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.001	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.001	0.001	0.1	0.1	0.65	0.08	0.50	0.15	0.17	0.02	0.12	0.05	0.46	0.14	0.31	0.10	0.52	0.02	0.39	0.17
		0.1	0	0.69	0.08	0.51	0.16	0.26	0.06	0.16	0.13	0.52	0.15	0.32	0.11	0.56	0.03	0.43	0.30
		0.01	0.01	0.19	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.09	0.00	0.02	0.00
		0.01	0	0.21	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.12	0.00	0.03	0.00
		0.001	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.001	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.001	0.0001	0.1	0.1	0.69	0.03	0.64	0.03	0.19	0.06	0.19	0.07	0.46	0.02	0.43	0.02	0.57	0.03	0.55	0.03
		0.1	0	0.72	0.03	0.67	0.13	0.30	0.11	0.28	0.12	0.49	0.02	0.45	0.02	0.64	0.04	0.60	0.16
		0.01	0.01	0.40	0.01	0.37	0.01	0.00	0.00	0.00	0.00	0.19	0.00	0.17	0.00	0.27	0.00	0.25	0.00
		0.01	0	0.43	0.01	0.39	0.01	0.01	0.00	0.00	0.00	0.23	0.00	0.20	0.00	0.31	0.00	0.29	0.00
		0.001	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.001	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.0001	0.001	0.1	0.1	0.67	0.09	0.50	0.30	0.19	0.06	0.14	0.05	0.44	0.28	0.32	0.21	0.57	0.03	0.42	0.20
		0.1	0	0.68	0.08	0.51	0.35	0.30	0.11	0.20	0.16	0.47	0.29	0.33	0.24	0.64	0.04	0.46	0.31
		0.01	0.01	0.31	0.00	0.13	0.11	0.00	0.00	0.00	0.00	0.11	0.11	0.01	0.01	0.27	0.00	0.07	0.02
		0.01	0	0.33	0.00	0.14	0.13	0.01	0.00	0.00	0.00	0.14	0.11	0.03	0.03	0.31	0.00	0.08	0.07
		0.001	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.001	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.0001	0.0001	0.1	0.1	0.76	0.11	0.52	0.14	0.25	0.02	0.23	0.06	0.53	0.13	0.46	0.10	0.63	0.06	0.59	0.26
		0.1	0	0.80	0.00	0.70	0.18	0.38	0.10	0.34	0.19	0.62	0.17	0.48	0.15	0.71	0.06	0.64	0.34
		0.01	0.01	0.65	0.08	0.49	0.15	0.17	0.02	0.11	0.05	0.46	0.14	0.32	0.11	0.51	0.03	0.41	0.17
		0.01	0	0.68	0.07	0.49	0.17	0.26	0.07	0.17	0.12	0.52	0.15	0.32	0.11	0.57	0.03	0.42	0.30
		0.001	0.001	0.20	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.10	0.00	0.02	0.00
		0.001	0	0.21	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.12	0.00	0.03	0.00

RESULTS

Selection models: As shown by THOMSON (1977), alleles at two neutral loci can increase in frequency due to linkage with a selected locus (hitchhiking), and a high disequilibrium may be created between these two neutral loci, even when they are initially in linkage equilibrium. Using this same model and examining normalized disequilibrium measures, we hypothesized that the disequilibrium between the two unselected loci would be closer to its minimum possible in the three-locus (more constrained) system than in the two-locus (less constrained) system, reflecting the fact that these loci are being "dragged" along on the selected haplotype.

This does seem to be the case generally when D' and D'' differ from each other. For the majority, and most likely, of the starting conditions, D'_{uu} and D''_{uu} (D' and D'' for the two unselected loci) begin at zero, while D'_{us} and D''_{us} (D' and D'' for an unselected with selected locus) begin at 1; however, while D'_{uu} can quickly reach values as high as those for D'_{us}, D''_{uu} does not increase as rapidly as does D'_{uu}. Specifically, for the two neutral loci the difference between D'_{uu} and D''_{uu} ($\delta_{uu} = |D'_{uu}| - |D''_{uu}|$) tends to be positive and

TABLE 5

Maximum δ_{uu} and δ_{us} values (in that order) reached with selection coefficients: $s_1 = 0.01$, $s_2 = 0$

		Selected locus															
R_1	R_2	*B* case 1		*C* case 2		*B* case 3		*C* case 4		*B* case 5		*C* case 6		*B* case 7		*C* case 8	
0.005	0.005	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.002	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.002	0.005	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.002	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.001	0.09	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
0.001	0.005	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.002	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
	0.001	0.21	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.12	0.00	0.03	0.00
	0.0001	0.43	0.01	0.39	0.01	0.01	0.00	0.00	0.00	0.23	0.00	0.20	0.00	0.31	0.00	0.29	0.00
	0.00001	0.47	0.00	0.49	0.00	0.03	0.00	0.00	0.00	0.26	0.00	0.26	0.00	0.37	0.00	0.38	0.00
0.0001	0.005	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.001	0.33	0.00	0.14	0.13	0.01	0.00	0.00	0.00	0.14	0.11	0.03	0.03	0.31	0.00	0.08	0.07
	0.0001	0.68	0.06	0.49	0.17	0.26	0.07	0.17	0.12	0.52	0.15	0.32	0.11	0.57	0.03	0.42	0.30
	0.00001	0.71	0.03	0.66	0.13	0.30	0.12	0.28	0.13	0.49	0.02	0.45	0.03	0.64	0.03	0.60	0.14
0.00001	0.001	0.26	0.00	0.14	0.16	0.03	0.00	0.00	0.00	0.12	0.13	0.03	0.05	0.37	0.00	0.10	0.06
	0.0001	0.69	0.08	0.50	0.35	0.30	0.12	0.16	0.16	0.47	0.29	0.33	0.24	0.64	0.03	0.43	0.31
	0.00001	0.80	0.10	0.70	0.20	0.38	0.10	0.33	0.19	0.63	0.19	0.48	0.15	0.71	0.06	0.63	0.35

greater in magnitude at any point in time than for the two δ_{us} values between the selected and a neutral locus ($\delta_{us} = |D'_{us}| - |D''_{us}|$). This latter value is often negative but only rarely positive. Note that in all cases considered the situation begins with δ_{uu} and δ_{us} values equal to zero. The δ values will also eventually return to a stable value of zero even before the disequilibrium itself has decayed to zero.

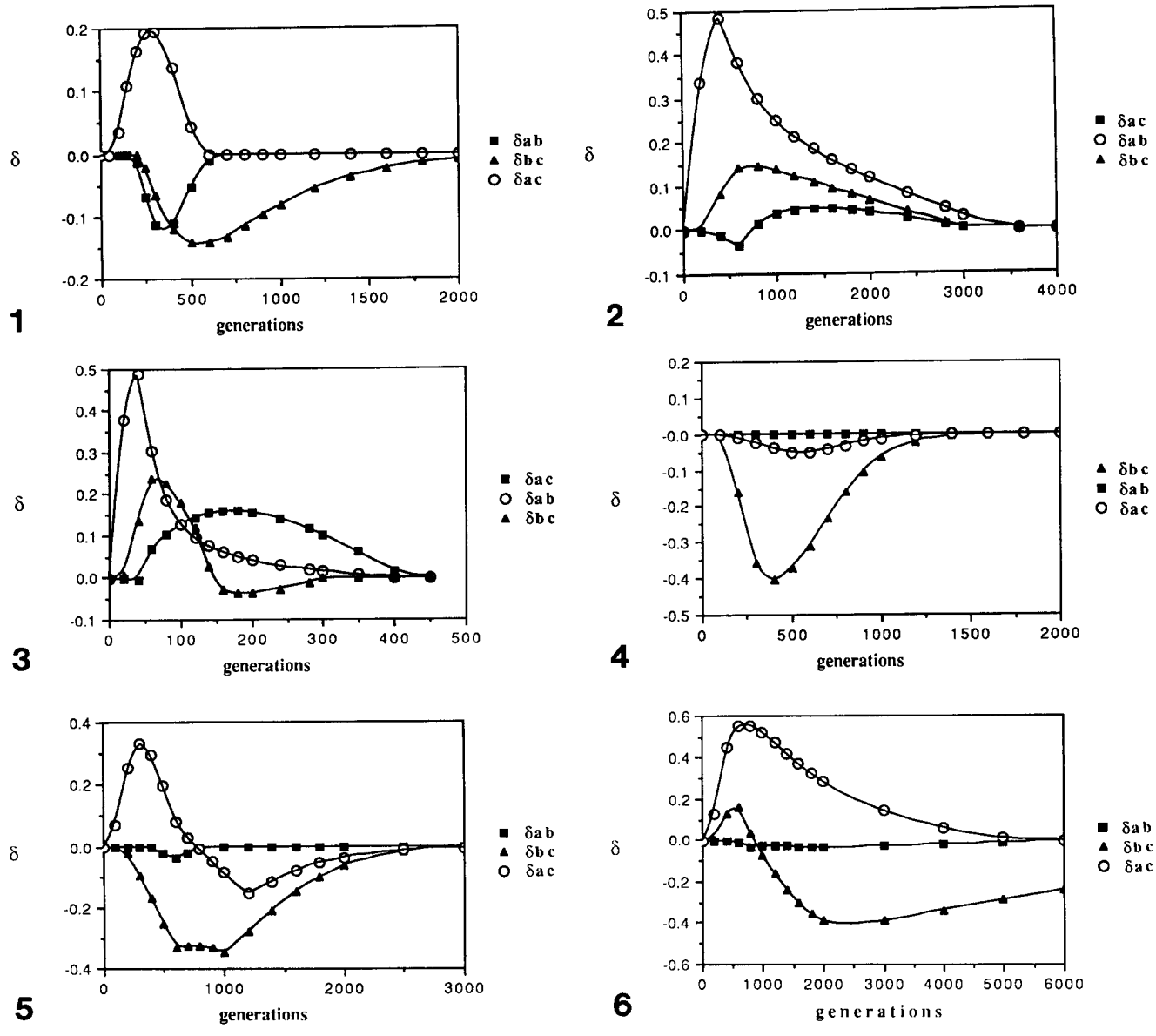
The δ value can be considered to be one measure of how the evolution of two loci is affected by additional nearby loci. When δ is not equal to zero, then the amount of disequilibrium has been affected not only by mutation at, and recombination between, these two loci, but by the evolutionary history of the third, linked locus.

Three basic patterns are observed under the selection models considered. In Tables 2 and 3, some examples of the patterns seen under various starting conditions are given. The most common pattern is $\delta_{uu} \geq 0 \geq \delta_{us}$ (designated pattern I), as is illustrated by Figure 1. This pattern is particularly common when the center locus (*B*) is the selected locus. δ_{uu} values have been observed as large as 0.80 (case 1; $s_1 = 0.1$, $s_2 = 0$, $R_1 = 0.0001$, $R_2 = 0.0001$, for example Table 4), and δ_{us} values can go as low as -0.52 (case 5; $s_1 = 0.1, s_2 = 0, R_1 = 0.0001, R_2 = 0.001$). The magnitudes of the δ values are, in general, correlated with the strength of selection. Directional selection ($s_1 > 0$, $s_2 = 0$) typically yields higher δ values than balancing selection ($s_1 = s_2$) given the same s_1 value (see Tables 3, 4 and 5).

The second most common pattern, where $\delta_{uu} > \delta_{us} \geq 0$, is designated pattern II (at least one δ_{us} value must be positive) (Figure 2). This pattern is most often seen when *C* (or by symmetry *A*) is the selected locus. In some cases, a pattern II occurs, but δ_{uu} will start to decrease as δ_{us} is still increasing, resulting in a period where $\delta_{us} > \delta_{uu} \geq 0$. This situation is designated pattern II′ (Figure 3). In our simulations a II′ pattern only occurred when *C* was the selected locus. The maximum positive value that δ_{us} reached in any of the simulation runs examined was 0.35. However, values for δ_{us} of this magnitude (both in II and II′ patterns) tended to only occur when $s_1 = 0.1$ (the maximum selection value examined) or when R_1 or $R_2 \leq 0.0001$ (see Tables 4 and 5). δ_{us} was never positive (pattern II or II′) when $s_1, s_2 \leq 0.001$, and rarely occurs if $R_1, R_2 \geq 0.001$ (and then only if $s_1 = 0.1$).

All δ values may also be negative (pattern III) (Figure 4). This pattern can occur with various starting conditions, but seems more frequent when $R_1 \neq R_2$. δ_{uu} does not normally reach values less than about -0.05. Combinations of these patterns may also occur (Figures 5 and 6).

In cases 10–15, where the system begins in linkage equilibrium and for cases 14–19, where a locus other than the rare allele (new mutant) is selected, all δ values always equal zero (data not shown) (this does not mean that the disequilibrium is zero, only that no difference exists between the three-locus (D'') and two-locus (D') normalized pairwise values). If selection is weak ($s_1, s_2 \leq 0.001$) or recombination is relatively high ($R = 0.01$), for any of the conditions examined, δ never deviates far from zero. As our



FIGURES 1–6.—Examples of the change in δ values with time (in generations) under various selection models. Note that δ_{uu} is always indicated by an open circle (O).

FIGURE 1.—An example of pattern I ($\delta_{uu} \geq 0 \geq \delta_{us}$): case 1, $s_1 = s_2 = 0.01$, $R_1 = R_2 = 0.001$. B is the selected locus such that $\delta_{uu} = \delta_{ac}$.

FIGURE 2.—An example of pattern II ($\delta_{uu} > \delta_{us} > 0$): case 2, $s_1 = s_2 = 0.01$, $R_1 = R_2 = 0.0001$. C is the selected locus such that $\delta_{uu} = \delta_{ab}$.

FIGURE 3.—An example of pattern II′ ($\delta_{us} > \delta_{uu} > 0$): case 2, $s_1 = 0.1$, $s_2 = 0$, $R_1 = 0.0005$, $R_2 = 0.001$. C is the selected locus such that $\delta_{uu} = \delta_{ab}$.

FIGURE 4.—An example of pattern III ($0 \geq \delta_{uu}, \delta_{us}$): case 7, $s_1 = 0.01$, $s_2 = 0$, $R_1 = 0.00001$, $R_2 = 0.005$. B is the selected locus such that $\delta_{uu} = \delta_{ac}$.

FIGURE 5.—An example of pattern I, III: case 1, $s_1 = 0.01$, $s_2 = 0$, $R_1 = 0.0001$, $R_2 = 0.001$. B is the selected locus such that $\delta_{uu} = \delta_{ac}$.

FIGURE 6.—An example of pattern II, I: case 5, $s_1 = 0.01$, $s_2 = 0$, $R_1 = 0.00005$, $R_2 = 0.0001$. B is the selected locus such that $\delta_{uu} = \delta_{ac}$.

model is a deterministic one, and ignores all effects of sampling and drift, we can only claim that in a very large, randomly mating population, large, nonzero δ values should only be created as a result of selection on a rare allele.

The two measures of three-way disequilibrium, D_{ABC} and D'_{ABC}, were also examined but do not show any patterns characteristic of selection. In the model presented here, the magnitude of the three-way disequilibrium begins at its maximum and generally decays with time (and recombination). The sign of the three-way disequilibrium may change from positive to negative and back several times as it decays.

In summary, a positive δ ($= |D'| - |D''|$) value for only one of the three pairwise combinations in a three-locus system is a distinguishing feature indicating selection has occurred on the locus not present in the positive pairwise δ. If more than one δ value is positive but one is much larger than the others, then again this indicates selection on the locus not included in the larger δ term. When all three are negative, or two are positive but close in value, it cannot be determined

TABLE 6

Distribution of D'_{AB} (Table 7) and δ_{AB} (Table 8) based on 1000 samples from seven specific populations

	A	B	C	D	E1	E2	F	G1	G2
D'_{AB} value									
<−0.95	717	368	0	331	144	0	636	1	0
−0.95 to −0.85	0	0	0	0	0	0	0	0	0
−0.85 to −0.75	0	0	0	0	3	0	0	0	0
−0.75 to −0.65	0	0	0	1	23	1	0	0	0
−0.65 to −0.55	0	4	0	4	62	2	0	0	0
−0.55 to −0.45	0	9	1	10	59	15	0	0	0
−0.45 to −0.35	0	21	0	27	67	66	0	1	0
−0.35 to −0.25	0	31	13	20	59	153	0	1	0
−0.25 to −0.15	0	53	84	47	69	227	0	5	0
−0.15 to −0.05	1	52	242	45	70	230	0	5	0
−0.05 to 0.05	26	177	352	186	108	224	303	15	0
0.05 to 0.15	110	184	219	211	115	74	0	106	0
0.15 to 0.25	59	66	76	75	81	7	6	169	79
0.25 to 0.35	54	27	13	35	55	1	17	250	686
0.35 to 0.45	5	7	0	5	31	0	2	228	232
0.45 to 0.55	20	1	0	2	29	0	23	141	3
0.55 to 0.65	2	0	0	1	11	0	1	55	0
0.65 to 0.75	1	0	0	0	2	0	0	15	0
0.75 to 0.85	0	0	0	0	1	0	0	6	0
0.85 to 0.95	0	0	0	0	0	0	0	1	0
>0.95	5	0	0	0	11	0	12	1	0
δ_{AB} values									
<−0.25	2	0	0	0	230	260	1	3	0
−0.25 to −0.20	0	0	0	0	8	130	0	0	0
−0.20 to −0.15	1	0	0	0	27	143	3	0	0
−0.15 to −0.10	0	2	0	6	28	121	0	3	0
−0.10 to −0.05	10	10	7	12	18	77	2	9	0
−0.05 to −0.01	32	67	132	82	20	42	0	31	0
−0.01 to 0	30	107	315	101	20	39	2	29	0
0	925	815	546	799	455	186	992	62	1
0 to 0.01	0	0	0	0	11	1	0	19	1
0.01 to 0.05	0	0	0	0	37	1	0	86	3
0.05 to 0.10	0	0	0	0	39	0	0	118	80
0.10 to 0.15	0	0	0	0	21	0	0	126	299
0.15 to 0.20	0	0	0	0	17	0	0	113	398
0.20 to 0.25	0	0	0	0	19	0	0	124	186
>0.25	0	0	0	0	49	0	0	277	32

Populations are as follows: A, $p_A = p_B = p_C = 0.05$, all disequilibria equal 0, sample size (N) = 100. B, $p_A = p_B = p_C = 0.1$, all disequilibria equal 0, sample size (N) = 100. C, $p_A = p_B = p_C = 0.5$, all disequilibria equal 0 sample size (N) = 100. *D*, $p_A = p_B = p_C = 0.1$, $D_{AB} = D_{AC} = D_{BC} = 0.05$ ($D' = 0.55$), $N = 100$. E1, $p_A = 0.05$, $p_B = 0.3$, $p_C = 0.01$, $D_{AB} = 0$, $D_{AC} = 0.0095$, $D'_{AC} = 1.0$, $D_{BC} = 0.007$, $D'_{AB} = 1.0$, $D_{ABC} = 0.00665$, $N = 100$; E2, $N = 500$; F, $p_A = 0.05$, $p_B = 0.3$, $p_C = 0.01$, $D_{AB} = D_{AC} = D_{BC} = D_{ABC} = 0$, $N = 100$; G1, $p_A = 0.138$, $p_B = 0.160$, $p_C = 0.100$, $D_{AB} = 0.036$, $D'_{AB} = 0.31$, $D_{AC} = 0.057$, $D'_{AC} = 0.66$, $D_{BC} = 0.057$, $D'_{BC} = 0.68$, $N = 100$; G2, $N = 1000$. In this last case (G), the initial $\delta_{AB} = 0.16$ and corresponds to the observed values for the A1-B8-DR3 haplotype (with *A* and *B* representing A1 and DR3).

which is more likely to be the selected locus. Using the criteria of a large positive δ value as a test to identify a selected allele will be very conservative; only certain combinations of initial allele frequencies, selection coefficients, recombination values, and generations of selection will give a pattern where it is clear which is the selected locus.

Sampling error: The effect of sampling error is examined by taking 1000 random samples from populations with a given set of allele frequencies and disequilibrium values. Some representative examples are given in Table 6. In cases A, B and C ($p_A = p_B = p_C$) the population is in linkage equilibrium but sampling only 100 individuals can create a large amount of disequilibrium as measured by D'. The distribution of D' is generally symmetrical around zero, excluding D' values equal to −1 due to loss of one allele, whereas the distribution of δ is quite assymetrical with all values falling at or below zero. A similar pattern is seen for case D where the sampled population shows moderate allele frequencies ($p_A = p_B = p_C = 0.1$) and large disequilibrium ($D = 0.05$, $D' = 0.55$ for all combinations). When one allele is rare and in strong disequilibrium with alleles at the other two loci, positive δ values are observed (for example see case E which is equivalent to the starting conditions for cases 1 and 2 of the hitchhiking simulations). However, the frequency of positive δ values decreases rapidly as sample size is increased. The presence of a rare allele in the absence of disequilibrium does not lead to positive δ

values (case F). It is interesting to note that when sampling from a population with allele frequencies and disequilibria similar to that observed for the HLA haplotype A1-B8-DR3 (case G), the mean D'_{AB} and δ_{AB} (A and B representing A1 and DR3, respectively) do not tend to change from the starting values. Thus, small sample size does not seem on average to reduce the value of δ once it is positive.

Thus positive δ values seem to be created by sampling only in a situation similar to that which arises when a new mutant is rare and in strong disequilibrium with linked alleles. One could expect that under genetic drift, which is sampling effects repeated over many generations, one would most often see positive δ values when a new mutant arises, population size is relatively small, and recombination does not occur quickly enough to reduce disequilibria and non-zero δ values created.

Genetic drift: Genetic drift was examined using the starting conditions of cases 1 through 9 (see Table 1). The evaluation of these results is difficult as frequently the new mutant is lost and even when nonzero δ values appear they are frequently associated with rare alleles (less than 0.01) or cases where several haplotypes are missing from the population (*i.e.*, $D' = 1$ or -1). This is especially true for cases with low recombination ($R < 0.001$), where nonzero δ values and $D' = \pm 1$ were frequently observed. As almost any combination of two locus haplotype frequencies can reasonably occur under neutrality for $4Nc < 10$ (HUDSON 1985), which corresponds in this case to $R < 0.0005$, analysis of δ values in this range (*i.e.*, when $4Nc < 10$) may not prove meaningful either.

For cases 3 and 4, positive (defined as values greater than 0.005) δ values were never observed when the recombination between the two loci was 0.001 or greater. For cases 1, 2, 5, 6, 7, 8 and 9, positive δ values were never observed when the recombination between the two loci was 0.005 or greater. Negative δ values were observed under all cases for values of $R \leq 0.005$, but were generally of small magnitude for $R > 0.001$. Some results of drift on the D' and δ values for cases 1 and 2 using various recombination values are given in Tables 7 and 8.

As positive δ values were observed for $R < 0.005$, it is of interest to determine the frequency and magnitude of positive δ values. Most of the observed positive values were quite small and $\delta \geq 0.1$ was infrequent. For cases 1, 2, 5, 6, 7, 8 and 9, the frequency of positive δ values for any of the three pairwise combinations among those runs which still maintained two alleles at all three loci was, for the first 200 generations, approximately 1–3% when $R1$ and R_2 were equal to 0.002 and was 5–10% when R_1 and R_2 were equal to 0.001. The frequency of positive δ values tends to decrease after 200 generations and is low (less than 1%) after 500 generations for $R \geq 0.001$. This is apparently because the disequilibrium decreases over time and because sampling has less effect as the rare allele increases in frequency in the population.

The frequency of positive δ values greater than 0.1 can also be considered. From 0 to 200 generations, the frequency of values this high never exceeded 1% or 4%, for recombination values of 0.002 or 0.001, respectively, for any of the cases. After 200 generations, positive values this large were rarely seen (less than 1%).

Although this analysis only covers a few specific conditions, it appears that genetic drift is unlikely to create positive δ values when $R_1, R_2 \geq 0.005$ and will do so only rarely when $R_1, R_2 > 0.001$. A conservative estimate of the effective human population size was used ($N_e = 5000$) so if the value of $N_e = 10{,}000$ that has been estimated for Caucasians of European descent is correct, then genetic drift would only create positive δ values for even smaller recombination values.

Admixture: Allele frequencies in two populations under admixture need to be quite different to create large amounts of disequilibrium in the admixed population. The results here indicate that even if two-locus disequilibrium is observed by such admixture, it proved to be almost impossible to create a situation where the δ values are also nonzero.

In Table 9, some examples are given where disequilibrium is nonzero, and the D' values are large but the δ values are zero. For example, when the frequencies of the p_A, p_B, and p_C alleles are all 0.2 in population-1 and 0.9 in population-2, then large amounts of disequilibrium are generated ($D' = 0.5$ for all pairwise combinations) but the δ values are all equal to zero. Only when the differences in these allele frequencies were even more extreme were nonzero δ values generated. No conditions could be found where any of the three δ values were negative. Admixture between populations with such extreme differences in allele frequency, would be unlikely to occur without some suspicion on the part of the experimenter. In realistic situations, non-zero δ values are unlikely to be explained by admixture.

Application to HLA data: The HLA region loci examined from the Provinces Francaise study (CAMBON-THOMSEN and OHAYON 1986; CAMBON-THOMSEN *et al.* 1989) consist of three class I loci (*A*, *B*, and *C*), one class II locus (*DR*) and three class III loci (*C4A*, *C4B*, and *Bf*) (*C4A* and *C4B* are involved in the classical and *Bf* in the alternate complement pathways); their relative positions and recombination distances are given in Figure 7. The fourteen most common haplotypes (with frequencies greater than

TABLE 7

Distribution D values under genetic drift sampled at 100 generations for 500 or 1000 repetitions (only cases where all six alleles still existed in the population were considered)

Case (see Table 1):	1	1	1	1	2	2	2	2
R_1,R_2:	0.0001	0.001	0.002	0.005	0.0001	0.001	0.002	0.005
Repetitions	500	1000	1000	500	500	1000	1000	500
	D'_{AC}	D'_{AC}	D'_{AC}	D'_{AC}	D'_{AB}	D'_{AB}	D'_{AB}	D'_{AB}
<−0.95	5	5	3	0	8	8	6	0
−0.95 to −0.85	9	10	3	1	7	20	12	4
−0.85 to −0.75	16	10	14	2	12	32	27	6
−0.75 to −0.65	11	18	21	9	12	31	32	6
−0.65 to −0.55	17	30	48	8	16	36	33	23
−0.55 to −0.45	18	26	51	19	10	48	54	9
−0.45 to −0.35	29	29	66	29	21	61	37	22
−0.35 to −0.25	27	35	69	27	26	58	62	32
−0.25 to −0.15	21	30	66	43	29	58	64	38
−0.15 to −0.05	28	28	67	50	17	57	68	48
−0.05 to 0.05	53	34	118	83	38	82	108	73
0.05 to 0.15	52	63	116	90	50	118	124	47
0.15 to 0.25	51	46	83	34	45	92	82	25
0.25 to 0.35	35	24	78	12	22	77	71	10
0.35 to 0.45	24	18	30	10	23	27	34	11
0.45 to 0.55	17	10	19	0	18	28	24	1
0.55 to 0.65	10	5	8	2	6	22	10	2
0.65 to 0.75	2	2	3	1	4	13	7	0
0.75 to 0.85	3	2	0	0	4	5	5	0
0.85 to 0.95	1	0	0	0	2	2	0	0
>0.95	0	1	0	0	1	1	2	0

Only D'_{AC} values are reported for case 1 and D'_{AB} values for case 2, as the other pairwise combinations almost never yield positive δ values.

TABLE 8

Distribution δ values under genetic drift sampled at 100 generations for 500 or 1000 repetitions (only cases where all six alleles still existed in the population were considered)

Case (see Table 1):	1	1	1	1	2	2	2	2
R_1,R_2:	0.0001	0.001	0.002	0.005	0.0001	0.001	0.002	0.005
Repetitions	500	1000	1000	500	500	1000	1000	500
	δ_{AC}	δ_{AC}	δ_{AC}	δ_{AC}	δ_{AB}	δ_{AB}	δ_{AB}	δ_{AB}
<−0.25	77	81	15	0	74	56	6	0
−0.25 to −0.20	10	23	15	0	7	11	5	0
−0.20 to −0.15	21	33	18	0	19	31	17	1
−0.15 to −0.10	20	39	31	2	13	37	24	0
−0.10 to −0.05	17	54	46	2	16	48	19	1
−0.05 to −0.01	15	82	80	10	19	50	51	2
−0.01 to 0	31	111	173	64	12	72	103	17
0	107	313	438	342	102	384	621	408
0 to 0.01	34	40	15	0	25	27	6	1
0.01 to 0.05	21	27	12	0	25	25	5	0
0.05 to 0.10	19	17	9	0	11	16	4	0
0.10 to 0.15	16	14	5	0	12	10	0	0
0.15 to 0.20	10	9	3	0	7	2	0	0
0.20 to 0.25	9	5	1	0	13	2	1	0
>0.25	22	8	1	0	16	5	0	0
freq. $\delta > 0.01$	0.23	0.09	0.04	0.00	0.23	0.07	0.01	0.00
freq. $\delta > 0.1$	0.13	0.04	0.01	0.00	0.13	0.02	0.00	0.00

Only δ_{AC} values are reported for case 1 and δ_{AB} values for case 2, as the other pairwise combinations almost never yield positive values.

0.25%) were selected for study; these are listed in Table 10.

Combinations of three loci at a time were examined for each of the fourteen haplotypes. Nine of these haplotypes contained three locus combinations which yielded positive δ ($|D'| - |D''|$) values. These nine haplotypes and the three-locus combinations yielding nonzero δ values are given in Table 11, a–i. Most of the pairwise combinations of alleles for these haplotypes have disequilibrium values significantly different from zero using the chi square test statistic. Note that a δ value can only be nonzero if D is nonzero (because D' and D'' must both equal zero when D equals zero). However, D_{AC} and D_{BC} may still place constraints on

TABLE 9

Examples

p_{A1}	p_{B1}	p_{C1}	p_{A2}	p_{B2}	p_{C2}	D_{AB}	D'_{AB}	D''_{AB}	D_{AC}	D'_{AC}	D''_{AC}	D_{BC}	D'_{BC}	D''_{BC}	δ_{AB}	δ_{AC}	δ_{BC}
0.1	0.1	0.0	0.5	0.5	0.5	0.4	0.19	0.19	0.05	0.29	0.29	0.05	0.29	0.29	0	0	0
0.3	0.01	0.05	0.1	0.05	0.01	0.002	0.33	0.33	0.002	0.08	0.08	0.00	0.50	0.50	0	0	0
0.01	0.01	0.01	0.5	0.5	0.5	0.06	0.06	0.06	0.06	0.32	0.32	0.32	0.32	0.32	0	0	0
0.2	0.2	0.2	0.9	0.9	0.9	0.123	0.495	0.495	0.123	0.495	0.495	0.123	0.495	0.495	0	0	0
0.1	0.1	0.1	0.8	0.8	0.8	0.123	0.495	0.495	0.123	0.495	0.495	0.123	0.495	0.495	0	0	0
0.1	0.1	0.1	0.9	0.8	0.8	0.14	0.622	0.595	0.14	0.622	0.595	0.123	0.495	0.063	0.11	0.11	0.05
0.1	0.1	0.1	0.9	0.9	0.9	0.16	0.64	0.50	0.16	0.64	0.50	0.16	0.64	0.50	0.14	0.14	0.14

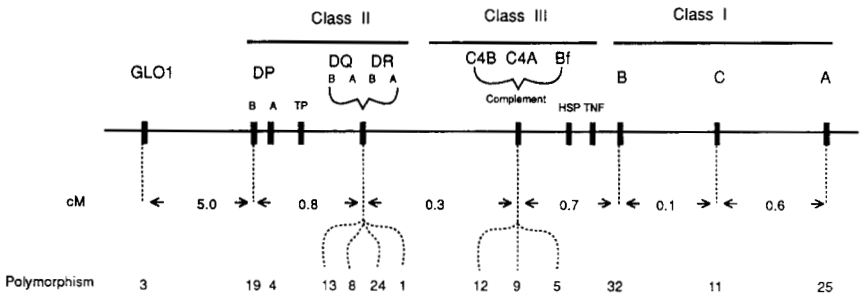


FIGURE 7.—Recombinational map of the HLA region indicating known number of alleles at each locus, as well as the polymorphism accounted for in the Provinces Francaises survey. Total allelic polymorphism is from the WHO Nomenclature Committee (1990). TNF, HSP and TP are the loci for tumor necrosis factor, heat shock protein, and transporter protein, respectively.

TABLE 10

Haplotypes examined from the "Provinces Francaises" study

Haplotype[a] at Locus:							Estimated frequency
A	*C*	*B*	*C4B*	*C4A*	*Bf*	*DR*	
1	7	8	1	Q0	S	3	0.0342
29	0	44	1	3	F	7	0.0212
3	7	7	1	3	F	2	0.0169
2	5	44	Q0	3	S	4	0.0121
	5	18	Q0	3	F1	3	0.0108
	4	35	1	3	S	11	0.0090
2	3	62			S	4	0.0089
1	6	57	1	6	S	7	0.0081
	6	13	1	3	S	7	0.0081
	3	60	1	3	S	4	0.0067
3	4	35	Q0		F	1	0.0060
	6	50	1	2	S07	7	0.0036
2	7	18	1	3	S	11	0.0032
2	0	51	1	3	S	2	0.0025

[a] For each haplotype, allele names are given under the locus name. A blank space indicates that no single allele at this locus is strongly associated with the remaining haplotype.

the value of D_{AB} even when one of them is equal to zero.

Results of Provinces Francaises data analysis: Of the 14 HLA haplotypes examined, nine showed positive δ values indicative of selection. Many three-way combinations consisted of δ values all equal to zero. As indicated in the results of our simulations, large δ values only occur when recombination is small ($R < 0.01$), selection is sufficiently strong ($s > 0.001$), and the selection occurs on a rare allele in strong disequilibrium with two other loci.

The total recombination distance covered by the HLA region is about 2.5% and recombination values between adjacent loci, when any combination of three are considered, range in this study from 0–1.7%. Between-loci recombination distances of about 0.1% ($R = 0.001$) or greater are optimal for the application of this method as this is when positive δ values are unlikely to be due to genetic drift. When recombination is much greater than 1%, nonzero δ values are not normally observed due to the rapid decay of disequilibrium. Of the loci examined here, recombination values only between the *C4A*, *C4B*, and *Bf* loci are below this optimal range (with no known recombination between pairs of these three loci).

Using the constrained disequilibrium values, alleles on nine haplotypes showed evidence of selection and are presented in Table 11, a–i. Most of the positive δ values occurred at, or near, the HLA-B locus. However two of these haplotype patterns indicated selection at the *Bf* locus (Tables 11, g and h), and one at the *DR* locus (Table 11i). For the A1-C7-B8-C4B1-C4AQO-BfS-DR3 haplotype (Table 11a), only when *B8* is one of the three alleles do strongly positive δ values occur between the other two alleles. A high δ (>0.1) occurs for a number of combinations and indicates that selection on this haplotype is most likely to have occurred on, or in tight linkage with, the *B8* allele. This haplotype is one of the most common haplotypes found in populations of Northern European origin and it also exhibits some of the largest disequilibrium values.

The A1-C6-B57-C4B1-C4A6-BfS-DR7 haplotype (Table 11b) shows positive δ values for both the *B* locus and the *C4A* locus. A very high δ value of +0.529 is shown for *B57* when paired against *C6* and *C4A6*. This observation leads one to conclude that *B57* is more likely to be the selected locus than is *C4A6*; however, the presence of positive values for two different alleles may indicate that one or more sites in

TABLE 11

δ values for three-way combinations of alleles at French HLA haplotypes

a. Haplotype[a]

A1	C7	B8	C4B-1	C4A-Q0	Bf-S	DR3
−0.024	−0.079	**0.125**				
−0.184		**0.133**		−0.184		
−0.124		**0.159**				−0.141
	−0.122	**0.101**		−0.015		
	−0.064	**0.096**				−0.026
		0.013		−0.069		0.0
		0.051			0.0	0.0
		0.109		0.0		−0.060

b. Haplotype

A1	C6	B57	C4B-1	C4A-6	Bf-S	DR7
0.0	−0.163	0.0				
−0.039		−0.065		−0.165		
−0.039	0.0					−0.074
0.0		−0.024				−0.050
	0.0	−0.064	0.0			
	−0.299	**0.529**		0.011		
	0.0	**0.071**			0.0	
		0.058			0.0	0.0
		0.030		**0.142**		−0.051

c. Haplotype

A3	C4	B35	C4B-Q0	Bf-S	DR1
0.0	−0.345	−0.073			
	0.045	**0.275**		0.0	
	0.0	−0.074			0.0

d. Haplotype

A29	C0	B44	C4B-1	C4A-3	Bf-F	DR7
		0.013		−0.069		0.0

e. Haplotype

A3	C7	B7	C4B-1	C4A-3	Bf-F	DR2
−0.010	−0.284	0.0				
−0.091		−0.007				−0.141
	−0.208	**0.048**				0.0

f. Haplotype

A30	C6	B13	C4B-1	C4A-3	Bf-S	DR7
0.0	−0.329	0.0				
0.0		0.0				−0.091
	0.0	**0.057**		0.0		
	0.0	**0.023**			0.0	
		0.034			0.0	0.0

g. Haplotype

A30	C5	B18	C4B-Q0	C4A-3	Bf-F1	DR3
0.0		−0.244			**0.004**	
	0.0	−0.045			**0.077**	
		−0.022	0.0		**0.080**	
		0.0			**0.187**	−0.061
				0.0	**0.093**	0.0

h. Haplotype

A2	C6	B50	C4B-1	C4A-2	Bf-S07	DR7
	−0.054	**0.048**			**0.344**	
	0.0			−0.096	**0.028**	
		−0.053		−0.124	**0.460**	
		−0.173			**0.161**	−0.070

i. Haplotype

A2	C5	B44	C4B-Q0	C4A-3	Bf-S	DR4
	−0.003	−0.087				0.0
	0.0				−0.029	**0.006**
		0.0			−0.033	**0.032**

[a] The alleles at three loci of the given haplotype are examined at one time. Pairwise Δ values are reported under the "constraining" locus, *i.e.*, the locus that imposes additional constraints (*D″*) on the pairwise disequilibrium value between the two other loci. For example, when considering the trio A1-C7-B8, the Δ for A1-C7 is reported under B8. When a large positive Δ value between two of three loci is observed, it usually indicates that the "constraining" locus is being selected. As there are 35 three-way combinations for each locus, only combinations yielding a Δ value with a positive or negative magnitude of 0.05 or greater were reported.

the region of these loci may be selected. At least 19 gene loci are known to be located between the *B* and *C2* loci (SPIESS, BRESNAHAN and STOMINGER 1989), including tumor necrosis factor and a heat shock protein gene.

The A3-C4-B35-C4A90-BfS-DR1 haplotype (Table 11c) shows only one strongly positive value, but this value is high ($\delta = 0.275$) and again it is the *B* locus that is the most likely site for being a selected allele. Slightly positive δ values are also observed for *B44* on the A29-C0-B44-C4B1-C4A3-BfF-DR7 haplotype (Table 11), *B7* on the A3-C7-B7-C4B1-C4A3-BfF-DR2 haplotype (Table 11e), and *B13* on the A30-C6-B13-C4B1-C4A3-C4A3-BfS-DR7 haplotype (Table 11f). Many of these values are small and do not provide strong evidence for selection, however it is of interest that the *B* locus seems to show positive δ values frequently.

The *Bf* locus has two common alleles, *S* and *F*, which together have a frequency of 97.4% in the French population (CAMBON-THOMSEN and OHAYON 1986) and two rare alleles *F1, SO7.* (There is a third rare allele, *Bf-V*, which is extremely rare with a frequency of only 0.001 and is not considered in this analysis.) One does not in general expect very common alleles to show disequilibrium patterns indicative of selection because they are presumably older and recombination has had time to obscure these patterns. Interestingly, both rare alleles at the *Bf* locus, *F1* and *SO7*, show δ patterns with very high positive values (particularly Bf-SO7) consistent with selection at this locus (Table 11, g and h).

A positive δ value is also observed for *DR4* on the

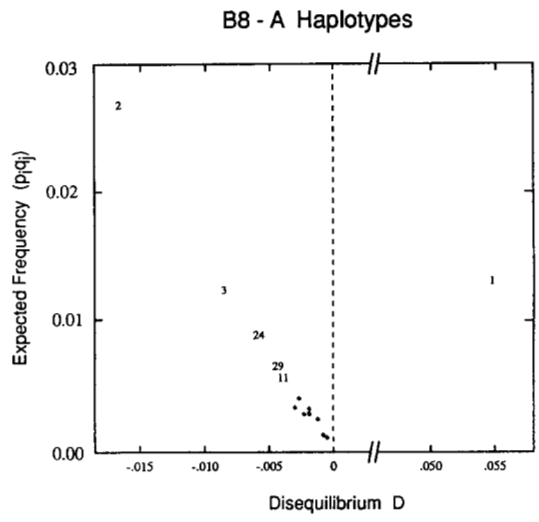


FIGURE 8.—All A-B haplotypes containing the alleles *B8*, where numbers indicate the allele designation at the *A* locus in the disequilibrium space, defined by *D*, the classical linkage disequilibrium measure.

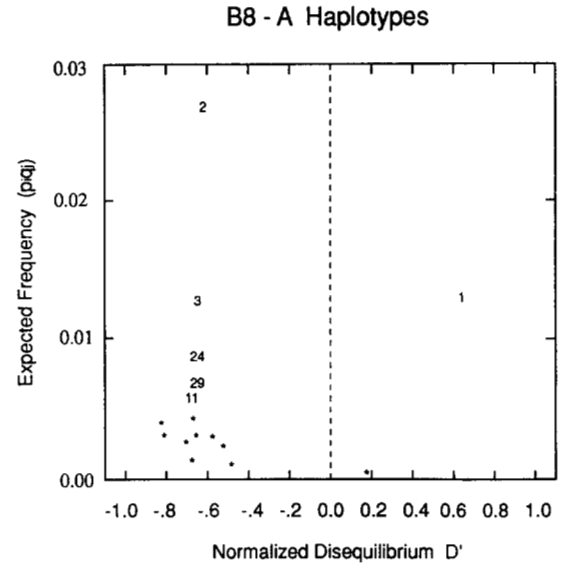


FIGURE 9.—All A-B haplotypes containing the alleles *B8*, where numbers indicate the allele designation at the *A* locus in the disequilibrium space, defined by the normalized disequilibrium measure *D'* and by the expected haplotype frequency under random association (the product of each of the constituent alleles). This pattern of haplotype, with one (or a few) frequent haplotypes in the positive space (here A1-B8), and all other haplotypes clustered around the single negative *D'* value is indicative of selection. Rare haplotypes are indicated with an asterisk.

A2-C5-B44-C4BQ0-C4A3-BfS-DR4 haplotype (Table 11i), however this value is small.

Application of linkage disequilibrium pattern analysis to the Provinces Francaises data identified six of the above haplotypes as showing patterns strongly suggestive of selection: A1-C7-B8-DR3, A3-C7-B7-DR2, A29-C0-B44-DR7, A1-C6-B17-DR7, C6-B13-DR7, and A3-C4-B35 (in decreasing order of strength). These correspond to the first six haplotypes listed in Table 11, a–f. The class III loci are not amenable to linkage disequilibrium pattern analysis, given their relatively low level of polymorphism. Thus it is not surprising that the two haplotypes involving rare *Bf* alleles (Tables 11, g and h) were not identified by this method. The DR4 haplotype (Table 8i) showed weak evidence for selection. (Allelic subdivision of DR4 using allele specific oligonucleotide probes gives stronger patterns of linkage disequilibrium, increasing the likelihood of detection of selection events in the future) (BEGOVICH, ERLICH and KLITZ 1991).

The strongest pattern indicative of selection is for the *B8* allele with the *A* locus alleles (Figures 8 and 9). The haplotype A1-B8 has an observed frequency of 0.068, and expected frequency of 0.0128 under random association, and a *D'* of +0.64. The only other B8 haplotype having positive disequilibrium is the rare A10n (for "10 new," which includes the non-A25 and non-A26 HLA-A10 variants). The rest of the B8 haplotypes occupy the negative space, with a linear relationship of the frequency of *D* with the frequency of the shared *A* allele (*i.e.*, non-*A1*, *A* locus alleles) (Figure 8), and have a *D'* value close to −0.64 (Figure 9), as expected under selection. Even with the large sample size of the Provinces Francaises data set, sampling error becomes evident in the rarer haplotypes. The other two point haplotypes (A with DR, and B

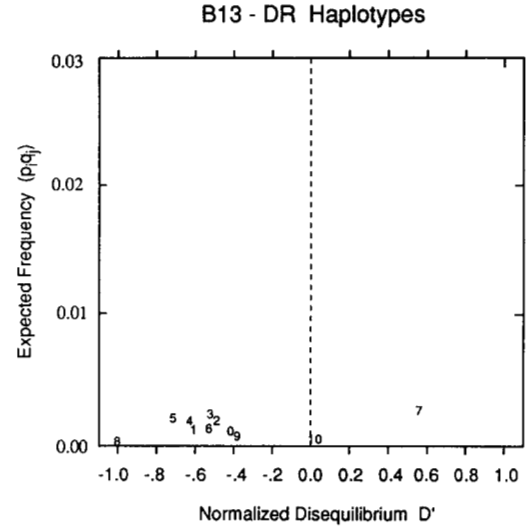


FIGURE 10.—The B-DR haplotypes containing the allele *B13* in the disequilibrium space, defined by *D'*. Despite its rareness, B13-DR7 appears to have been selected.

with DR) comprising the A1-B8-DR3 haplotype show similar, slightly less striking, patterns as that seen for the B8-A haplotypes (data not shown).

Rare alleles exhibit selection patterns as well. The allele *B13* with a frequency of only 0.016 is a case in point (Figure 10). The haplotype B13-DR7 is alone in the positive space with the other DR haplotypes clustered around $D' = -0.55$.

For contrast to cases indicating selection, we include the A2-B haplotypes (Figure 11), as an example illus-

trative of the great majority of cases, where the distinctive pattern indicative of recent selection is not present. In this case the D' values are scattered throughout the positive and negative spaces.

Although only a few haplotype combinations show strongly nonzero δ values, many haplotypes that have undergone selection and have large disequilibrium values may not show nonzero δ values, for various reasons. For the sample cases examined in this study, δ values did not deviate much from zero when selection was weak ($s \leq 0.001$) or even with moderate selection ($s_1 = 0.01$) under certain starting conditions (when the allele(s) at the locus or loci closest to the "new mutant" began at a high frequency, *i.e.*, cases 3, 4 and 6, and cases where selection occurred only after alleles were at equilibrium) (see Tables 2–5). A small δ value does not necessarily mean that there is not a large amount of disequilibrium present; a δ value may begin large but may return to zero long before the disequilibrium has decayed to zero. δ values tended to zero after several hundred generations for $s = 0.1$ or several thousand generations for $s = 0.01$. Subsequent selection events on other loci would also conceivably obscure older δ patterns. Nonzero δ values seem to be most likely observed when reasonably strong selection has acted on a recent mutant.

The main purpose of the constrained disequilibrium values is not to identify selected haplotypes *per se*, but to identify which of several alleles on a putatively selected haplotype (based on evidence of strong disequilibrium values and not-rare allele frequencies) is the most likely to have been selected. This approach is expected to be conservative so that only a fraction of such events will be identified.

DISCUSSION

Although it is generally accepted that strong selective forces are acting on the HLA region, and it is even possible to identify which regions and amino acids of HLA class I and II loci seem to have been selected (HUGHES and NEI 1988, 1989; HEDRICK et al. 1991; HEDRICK, WHITTAM and PARHAM 1991), it has not been possible to directly identify the individual alleles that have been most strongly and/or most recently selected. The appeal of the present analysis using constrained disequilibrium values, combined with the method of disequilibrium pattern analysis, is the identification not only of selected haplotypes in the HLA region, but the particular allele on the haplotype which has most likely been selected. We can now look for amino-acid differences which distinguish selected alleles from their putative progenitor alleles.

Although loci in the HLA region are generally highly polymorphic, the *Bf* locus is the least polymorphic of the seven loci examined. The *B* locus, on the other hand, is the most polymorphic with 50 serologically defined alleles (and more at the sequence level). There is previous evidence that the class I and class II loci show allele-frequency distributions which are more even than neutral expectations, thus confirming that some form of balancing selection has occurred here. On the other hand, the class III loci allele-frequency distributions either fit neutrality expectations (*Bf, C4A, C4B*) or are more skewed than expected (*C2*) (KLITZ, THOMSON and BAUR 1986), suggesting that selection at the class III loci, if it occurs, is either weak or directional. The observations that the distribution of allele frequencies at the *Bf* locus does not differ significantly from neutrality expectations, and our evidence for selection acting on the rare alleles *F1* and *SO7* are not inconsistent. Both methods are conservative and will not detect all instances of selection.

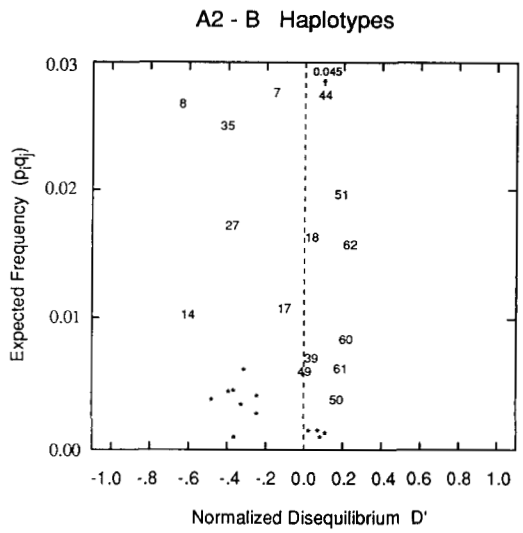


FIGURE 11.—The A-B haplotypes containing the allele *A2* in the disequilibrium space, defined by D'. Numerous haplotypes, in both the positive and negative space suggest the absence of recent selection favoring *A2*.

The evidence from this study supports the case that selection has acted on class I, class II and class III loci, in Northern European Caucasians. Selection has certainly occurred at other HLA loci besides HLA-B and Bf, but perhaps it has not been as strong or as recent. The high polymorphism at the *B* locus, the synonymous versus non-synonymous amino acid substitution rates, and lack of any evidence for a high mutation rate, suggests that a great amount of balancing selection has occurred (and may still be occurring) here, with only a small fraction (perhaps the most recent) of these selection events detectable by examining constrained disequilibrium values. There may also have been recent directional selection on a few alleles, in response to a new pathogen for example, on top of the general underlying balancing selection.

Additional selection events may be hidden in serologically defined HLA specificities, which actually

consist of two or more alleles. DR4, for example, has been split into eight new subtypes (WHO Nomenclature Committee 1990) using molecular methods. The application of allele specific oligonucleotide probes to type for 30 DRβ1 alleles on 268 Caucasian haplotypes suggested selection on the two common DR4 splits based on disequilibrium pattern analysis (BEGOVICH, ERLICH and KLITZ 1991). Analysis of disequilibrium values for additional class II loci (*DQA, DQB* and *DPB*) typed with allele specific probes may reveal additional selection on class II loci.

The French population sample used here is large, so the results observed are probably not dominated by sampling error. Admixture is also unlikely to create the patterns we observe. There is as yet no statistical test for determining which are significant positive or significant negative values for δ. The expectation and variance of two-locus disequilibrium values have been examined by extensive simulation (HUDSON 1985; HEDRICK and THOMSON 1986) and are highly dependent on the population size, number of alleles per locus, and sample size, and thus, a general statistical test is not possible. Examining the distribution patterns of two or three-locus disequilibrium measures may not provide "proof" that selection has taken place, but is practical to apply and provides important clues as to which specific alleles (and their neighborhood on the chromosome) may be worthy of more detailed analysis.

The extraordinary HLA diversity, age of alleles and patterns of linkage disequilibrium are apparently the result of strong selection operating on moderate rates of mutation, recombination and conversion-generated new variation (see *e.g.*, KUHNER *et al.* 1990, 1991). The trans-specific polymorphism (KLEIN 1987) of HLA variation cannot be explained by neutral mutations since they could not be maintained in populations for this length of time. Selection to maintain diversity of class I and class II antigen recognition sites is clearly occurring (HUGHES and NEI 1988, 1989), and balancing selection is a strong candidate to explain this diversity (TAKAHATA and NEI 1990). The frequently reported associations of HLA antigens with various disease susceptibilities makes it plausible that exposure to new pathogens may, in addition, occasionally result in positive selection for a resistant allele(s) or negative selection against a susceptible allele(s). The ability to distinguish specific selected alleles does not by itself distinguish among different selective mechanisms. However, the knowledge of which amino acid sites are of most importance combined with sequence comparisons between an allele showing signs of selection and its closest relatives, may indicate what amino acid site change was most likely to have given the selective advantage to a particular allele.

The difficulty of detecting selection events has long been recognized, and is reflected in the literature by the paucity of examples of selection in natural populations. The need to examine data in as many complementary ways as possible to detect clues of past and present selection events is thus of paramount importance. We introduce the approach of examining pairwise disequilibrium values in the context of three locus systems as a useful example of exploratory data analysis, with the full knowledge that a statistical test is not available for the method. It is gratifying in our study that the results of application of an alternate method-disequilibrium pattern analysis, are in concordance with the results using the constrained disequilibrium values. The two methods test different features of the data, so such agreement should not always be expected. Any evidence provided on selection events allows for a more informed discussion on the evolution of this complex multi-gene family and the evolution of disease predisposing genes and epitopes in the region. The methodology is not restricted to HLA but applies to any polymorphic system of closely linked genes.

We thank our colleagues BRUCE RISKA, MARY KUHNER and LIANNE VOELM for their helpful comments on the manuscript. This project was supported by National Institutes of Health grant HD12731 (W. P. R., N. B., W. K. and G. T.).

LITERATURE CITED

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

BEGOVICH, A., H. A. ERLICH and W. KLITZ, 1991 Recombination, polymorphism and disequilibrium across the HLA class II region. *J. Immunol.* (in press).

CAMBON-THOMSEN, A., and E. OHAYON, 1986 Analyse des données génétiques sur l'echantillon global des Provinces Françaises, pp. 297–322 in *Human Population Genetics*, edited by E. OHAYON and A. CAMBON-THOMSEN. Inserm, Paris.

CAMBON-THOMSEN, A., N. BOROT, M. NEUGEBAUER, A. SEVIN and E. OHAYON, 1989 Inter-regional variability between 15 French provinces and Quebec. Collegium Anthropologicum **13:** 24–41.

DUPONT, B., 1989 *Immunobiology of HLA*, Vols. 1 and 2. Springer-Verlag, Berlin.

FIGUEROA, F., E. GUNTHER and J. KLEIN, 1988 MHC polymorphism predating speciation. Nature **335:** 265–267.

GYLLENSTEN, U. B., and H. A. ERLICH, 1989 Ancient roots for the polymorphism at the HLA-DQα locus in primates. Proc. Natl. Acad. Sci. USA **86:** 9986–9990.

GYLLENSTEN, U. B., D. LASHKARI and H. A. ERLICH, 1990 Allelic diversification at the class II DQβ locus of the mammalian major histocompatibility complex. Proc. Natl. Acad. Sci. USA **87:** 1835–1839.

HEDRICK, P., and G. THOMSON, 1986 A two-locus neutrality test: applications to humans, E. coli, and lodgepole pine. Genetics **112:** 135–156.

HEDRICK, P. W., T. S. WHITTAM and P. PARHAM, 1991 Heterozygosity at individual amino acid sites: extremely high levels for the HLA-A and -B genes. Proc. Natl. Acad. Sci. USA **88:** 5897–5901.

HEDRICK, P. W., W. KLITZ, W. P. ROBINSON, M. K. KUHNER and

G. THOMSON, 1991 Population genetics of HLA, pp. 248–271 in *Evolution at the Molecular Level*, edited by R. SELANDER, A. CLARK and T. WHITTAM. Sinauer, Sunderland, Mass.

HUDSON, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109:** 611–631.

HUGHES, A. L., and M. NEI, 1988 Pattern of nucleotide substitution at MHC-class I loci reveals overdominant selection. Nature **335:** 167–170.

HUGHES, A. L., and M. NEI, 1989 Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. USA **86:** 958–962.

KLEIN, J., 1987 Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. Hum. Immunol. **19:** 155–162.

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

KLITZ, W., G. THOMSON and M. P. BAUR, 1986 Contrasting evolutionary histories among tightly linked HLA loci. Am. J. Hum. Genet. **39:** 340–349.

KLITZ, W., G. THOMSON, N. BOROT and A. CAMBON-THOMSON, 1991 Evolutionary genetics of HLA. Evol. Biol. (in press).

KUHNER, M. K., S. WATTS, W. KLITZ, G. THOMSON and R. S. GOODENOW, 1990 Gene conversion in the evolution of both the H-2 and Qa class I genes of the major histocompatibility complex. Genetics **126:** 1115–1126.

KUHNER, M. K., D. A. LAWLOR, P. ENNIS and P. PARHAM, 1991 Gene conversion in the evolution of the human and chimpanzee MHC class I loci. Tissue Antigens (in press).

LAWLOR, D. A., F. E. WARD, P. D. ENNIS, A. P. JACKSON and P. PARHAM, 1988 HLA-A and -B polymorphism predate the divergence of the humans and chimpanzees. Nature **335:** 268–271.

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

MCKUSICK, V. A., 1988 *Medelian Inheritance in Man*, Ed. 8. Johns Hopkins University Press, Baltimore.

NEI, M., and D. GRAUR, 1984 Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. **17:** 73–113

NEUGEBAUER, M., J. WILLIAMS and M. P. BAUR, 1984 Analysis of multilocus pedigree data by computer, in *Histocompatibility Testing 1984*, edited by E. D. ALBERT, M. P. BAUR and W. R. MAYR. Springer-Verlag, Berlin.

ROBINSON, W. P., 1989 Population Genetic Analysis of Selection and Diseases Associations at HLA gene loci. Ph.D. thesis, University of California, Berkeley.

ROBINSON, W. P., M. A. ASMUSSEN and G. THOMSON, 1991 Three-locus systems impose additional constraints on pairwise disequilibria. Genetics **129:** 925–930.

SPIESS, T., M. BRESNAHAN and J. L. STOMINGER, 1989 Human major histocompatibility complex contains a minimum of 19 genes between the complement cluster and HLA-B. Proc. Natl. Acad. Sci. USA **86:** 8955–8958.

TAKAHATA, N., and M. NEI, 1990 Allelic geneology under overdominant and frequency dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124:** 967–978.

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

THOMSON, G., and M. BAUR, 1984 Third order linkage disequilibrium. Tissue Antigens **24:** 250–255.

THOMSON, G., and W. KLITZ, 1987 Disequilibrium pattern analysis. I. Theory. Genetics **116:** 623–632.

WEIR, B., and W. G. HILL, 1986 Nonuniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. **38:** 776–778.

WHO Nomenclature Committee, 1990 Nomenclature for factors of the HLA system, 1989. Immunogenetics **31:** 131–140.

Communicating editor: A. G. CLARK