A Method for Estimating the Intensity of Overdominant Selection From the Distribution of Allele Frequencies

Montgomery Slatkin and Christina A. Muirhead

Department of Integrative Biology, University of California, Berkeley, California 94720-3140

Manuscript received May 22, 2000
Accepted for publication August 24, 2000

ABSTRACT

A method is proposed for estimating the intensity of overdominant selection scaled by the effective population size, $S = 2Ns$, from allele frequencies. The method is based on the assumption that, with strong overdominant selection, allele frequencies are nearly at their deterministic equilibrium values and that, to a first approximation, deviations depend only on S. Simulations verify that reasonably accurate estimates of S can be obtained for realistic sample sizes. The method is applied to data from several loci in the major histocompatibility complex (*Mhc*) in numerous human populations. For alleles distinguished by both serological typing and the sequence of the peptide-binding region, our estimates of S are comparable to those obtained by analysis of DNA sequences in showing that selection is strongest on *HLA-B* and weaker on *HLA-A*, *HLA-DRB1*, and *HLA-DQA1*. The intensity of selection on *HLA-B* varied considerably among populations. Two populations, Native American and Inuit, showed an excess rather than a deficiency in homozygosity. Comparable estimates of S were obtained for alleles at *Mhc* class II loci distinguished by serological reactions (serotyping) and by differences in the amino acid sequences of the peptide-binding region (molecular typing). A comparison of two types of data for *DQA1* and *DRB1* showed that serotyping led to generally lower estimates of S.

THE extensive polymorphism at the major histocompatibility complex (*Mhc*) loci in humans and other vertebrates is probably attributable to balancing selection caused by the superior performance of the immune system of individuals more heterozygous at these loci. Hughes and Yeager (1998) summarize evidence in favor of this theory. In this article we describe a simple method for using allele frequencies at a locus to estimate the intensity of selection in favor of heterozygotes. Our method is based on the assumption that allele frequencies are held at almost the frequency expected under selection alone and that deviations from those frequencies are determined by the balance achieved between genetic drift and selection. In a previous article, we showed that making such an assumption leads to an excellent approximation for the numbers of alleles maintained at a locus under the combined effects of genetic drift, mutation, and overdominant selection (Slatkin and Muirhead 1999). We first develop the basic theory, then test the performance of our method on simulated data, and finally apply it to several published sets of allele frequencies at *Mhc* loci in humans. Our results are similar to those of Satta *et al.* (1994), who estimated selection intensities from the ratio of silent to replacement changes in the peptide-binding region of *Mhc* proteins. Edwards and Hedrick (1998) review other methods for estimating selection intensities in the *Mhc* region.

METHOD

Our method assumes a single locus with k alleles in a population containing N diploid individuals and symmetric overdominant selection in which every heterozygous individual has a fitness $1 + s$ relative to every homozygous individual. The frequency of the ith allele is x_i and, at the deterministic equilibrium under selection alone, $x_i = 1/k$. We assume that genetic drift results in small deviations of allele frequencies from $1/k$ and write the deviation at the ith allele as $\xi_i = x_i - 1/k$. One generation of selection and genetic drift changes the ξ_i to ξ_i'. We can use the standard theory of population genetics to compute the approximate mean and variance of $\Delta\xi_i = \xi_i - \xi_i'$ and the covariance between $\Delta\xi_i$ and $\Delta\xi_j$ for the ith and jth alleles ($i \neq j$) under the assumption that s is small and N is large,

$$E(\Delta\xi_i) \cong s\left(\frac{1}{k} + \xi_i\right)\left(F - \frac{1}{k} - \xi_i\right) \quad (1a)$$

$$\mathrm{Var}(\Delta\xi_i) \cong \frac{(1/k + \xi_i)\,(1 - 1/k - \xi_i)}{2N} \quad (1b)$$

$$\mathrm{Cov}(\Delta\xi_i, \Delta\xi_j) = -\frac{(1/k + \xi_i)\,(1/k + \xi_j)}{2N}, \quad (1c)$$

where $F = \Sigma_i x_i^2$ is the homozygosity.

Corresponding author: Montgomery Slatkin, Department of Integrative Biology, University of California, 3060 Valley Life Sciences Bldg., Berkeley, CA 94720-3140.
E-mail: slatkin@socrates.berkeley.edu

The general solution to a multidimensional diffusion equation based on (1) has been obtained by WATTERSON (1977) but it is difficult to use that result to derive an estimate of the intensity of selection. Here, we obtain approximate results for both the population and a sample from it by assuming that $F = 1/k$, the value at the equilibrium under selection, and that the ξ_t are much smaller than $1/k$. In other words, we assume that genetic drift results in only small deviations from the deterministic equilibrium. With these approximations

$$E(\Delta\xi_i) = -s\xi_i/k, \tag{2a}$$

$$\mathrm{Var}(\Delta\xi_i) = 1/(2Nk), \tag{2b}$$

$$\mathrm{Cov}(\Delta\xi_i, \Delta\xi_j) = -1/(2Nk^2), \tag{2c}$$

where (2b) assumes that $k \gg 1$, which is typically the case at *Mhc* loci. If k is small, then N in (2b) is divided by $1 - 1/k$. A diffusion equation for the joint probability distribution of $\xi_1, \ldots, \xi_k$ can be derived from (2). The solution to that equation tells us that the joint probability distribution of $\xi_1, \ldots, \xi_k$ is a symmetric multivariate normal with mean values of 0, variances $1/(4Ns)$, and covariances $-1/(4Nsk)$ (FELSENSTEIN 1977). Because $x_i = \xi_i + 1/k$, the joint distribution of $(x_1, \ldots, x_k)$ is multivariate normal with means $1/k$, with equal variances given by (2b) and with equal covariances given by (2c).

Under this model and under the assumption of strong selection, the joint distribution of allele frequencies in the population depends on only one combination of parameters, $4Ns$, which we denote by $2S$ to be consistent with previous notation. The problem is to estimate S from a sample from a population. A data set consists of a set of counts of each allele, $\{i_j\}$ $(j = 1, \ldots, k)$, where $\sum_{j=1}^{k} i_j = n$, the sample size ($n/2$ being the number of individuals sampled). In general, the true number of alleles at the locus is unknown so, in our analysis, k means the number of alleles found in the sample and the i_j are all nonzero. The value of k can and does differ among populations. Our method assumes in effect that differences among populations in the number of alleles found in a sample reflect differences among populations in the true number of alleles present. It would be easy to modify the method presented here so that the true number of alleles is assumed known and that some of the i_j are zero. Such a modification would increase the apparent variance in allele frequencies and reduce the estimated values of S.

To simplify the analysis, we further approximate the multivariate normal distribution of allele frequencies by a symmetric Dirichlet distribution,

$$\Pr(x_1, \ldots, x_k | a) = \frac{\Gamma(ka)}{(\Gamma(a))^k} \prod_{j=1}^{k} x_j^{a-1},$$

where the parameter a is chosen so that $S = k^2(1 + ka)/(k - 1)$, which ensures that both distributions have the same means, variances, and covariances. The reason for using a Dirichlet rather than a multivariate normal distribution is that the sampling distribution has a particularly simple form, the multinomial Dirichlet distribution,

$$\Pr(i_1, \ldots, i_k | a) = \frac{\prod_{j=1}^{k} (a)_{(i_j)}}{(ka)_{(n)}}, \tag{3}$$

where $(a)_{(i)} = a(a + 1) \ldots (a + i - 1)$ (JOHNSON *et al.* 1997). For large values of a, there is little difference between Dirichlet and multinomial distributions, and the use of a Dirichlet distribution ensures that all allele frequencies are positive.

Given the data, Equation 3 provides the likelihood of a and hence of S. We have written a short Mathematica (WOLFRAM 1996) program to find the maximum-likelihood estimate (MLE) of S, $\hat{S}$, and will distribute copies of that program upon request. This program either provides $\hat{S}$ or it indicates that the likelihood function is a monotonically increasing function of S, implying that there is no finite MLE value. In that case, we report the result as $\hat{S} = \infty$. The value of $\hat{S}$ is always positive because of the functional dependence of the likelihood on a.

Two extreme hypotheses, $S = 0$ and $S = \infty$, can be considered separately. Neutrality ($S = 0$) can be tested using either WATTERSON's (1977) test based on the observed homozygosity, F (also called the Ewens-Watterson test), or an exact test (SLATKIN 1994). We used the program described by SLATKIN (1996) to carry out the Ewens-Watterson test and provide the tail probability P_{H}. The value of P_{H} is the probability under neutrality ($S = 0$) of obtaining a value of F no larger than the observed value in a random sample of the same size containing the same number of alleles. A small value of P_{H} indicates too little homozygosity and can be the result of overdominant selection. A value of P_{H} close to 1 indicates excess homozygosity and can be the result of purifying selection. The exact test is similar but was not designed to be powerful against the specific alternative of overdominant selection.

The other extreme hypothesis is $S = \infty$, in which case the allele frequencies are at their deterministic equilibrium value, $1/k$, and the deviations from those frequencies are attributable to sampling only. The test for $S = \infty$ is then a test of the hypothesis that the sample is drawn from a multinomial distribution with equal probabilities and sample size n. For reasonably large sample sizes, a χ^2 test with $k - 1$ d.f. can be used and a probability of the data under the hypothesis that $S = \infty$, P_{M} (M for multinomial), can be obtained. A one-sided test is appropriate in this case because we assume random sampling is always present and the only question is whether there is significant deviation from equal frequencies. A value of $P_{\mathrm{M}} > 0.05$ indicates that the data do not reject the hypothesis that $S = \infty$. Our Mathematica program also computes P_{M}.

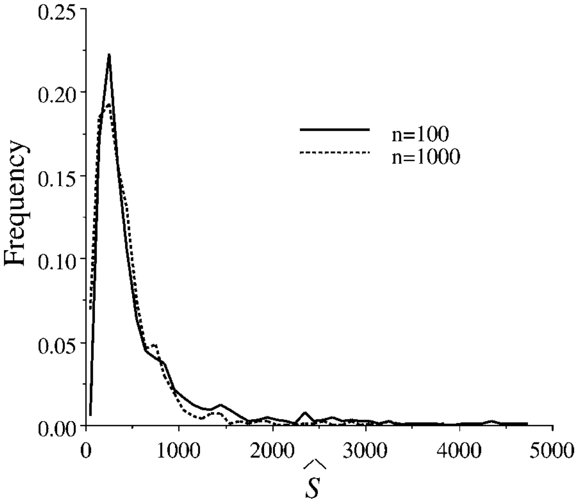

Figure 1.—Frequency distribution of $\hat{S}$ in two sets of 1000 replicates of a model with $N = 10{,}000$, $u = 10^{-6}$ ($M = 0.1$), and $s = 0.01$ ($S = 400$). For $n = 100$, $\hat{S}$ was finite in 902 replicates and for $n = 1000$, $\hat{S}$ was finite in 997 replicates (*cf.* Table 1).

For any data set, our program and the program to test neutrality provide three numbers, $\hat{S}$, P_M, and P_H. We found in some cases that $P_M < 0.05$ but $\hat{S} = \infty$, but we have found no cases in which $P_M > 0.05$ but $\hat{S}$ is finite. In the simulated data sets described in the next section we found several for which $P_M > 0.05$, but no such cases in any of the real data sets we analyzed.

Our method estimates S from the extent of variation in allele frequencies at a locus. The dependence on k is very weak. Our previous results (Slatkin and Muirhead 1999) and those of Takahata (1990) provide a formula relating the expected number of alleles to S and the scaled mutation rate, $M = Nu$:

$$k = \sqrt{\frac{2S}{\ln(S/32\pi M^2)}}. \tag{4}$$

Equation 4 can be solved for M as a function of k and S:

$$M = \sqrt{\frac{S}{32\pi}}e^{-S/k^2}. \tag{5}$$

When $\hat{S}$ is finite, our program calculates M from Equation 5.

SIMULATION TEST

We applied our method to simulated data generated by the program described in our earlier article (Slatkin and Muirhead 1999). For each replicate, the parameters are $S = 2Ns$, $M = Nu$, and n, the sample size. In all cases, we assumed $N = 10{,}000$. Samples were taken after the numbers of alleles reached stationarity. For each set of parameter values we generated 1000 replicate data sets.

Typical distributions of $\hat{S}$ are shown in Figure 1 and the summary of results is shown in Table 1. For the parameter values we used in our simulations, estimates of S are relatively unbiased although they have broad 95% confidence intervals. Averages and confidence intervals reported in Table 1 are based on all replicates

TABLE 1

Performance of method for finding $\hat{S}$, the MLE of $S = 2Ns$ in sets of 1000 replicate simulations for each set of parameter values

n	u	Average ($\hat{S}$)	95% C.I.	$\hat{S} < \infty$	$P_M > 0.05$
		A. $S = 400$ ($s = 0.02$)			
100	10^{-6}	598	(116, 2895)	902	383
1,000	10^{-6}	429	(74, 1419)	997	1
20,000	10^{-6}	372	(49, 1332)	1000	0
100	10^{-7}	527	(98, 2528)	778	530
1,000	10^{-7}	555	(104, 1996)	980	11
20,000	10^{-7}	543	(55, 1832)	997	0
		B. $S = 4000$ ($s = 0.2$)			
100	10^{-6}	5330	(1213, 20812)	720	796
1,000	10^{-6}	3139	(1162, 7831)	1000	1
20,000	10^{-6}	2532	(635, 5799)	1000	0
100	10^{-7}	4585	(917, 17174)	659	821
1,000	10^{-7}	3666	(1294, 10980)	998	2
20,000	10^{-7}	3229	(1018, 6769)	1000	0

$\hat{S} < \infty$ indicates the number of replicates in which finite estimates of $\hat{S}$ were obtained. $P_M > 0.05$ indicates the number of replicates in which the hypothesis of a multinomial sample with equal frequencies could not be rejected at the 5% level. In all cases the population size was $N = 10{,}000$. For $u = 10^{-6}$, each replicate was run for 100,000 generations and for $u = 10^{-7}$, 800,000 generations. The averages and 95% confidence intervals (C.I.) are based on only those replicates for which $\hat{S} < \infty$.

TABLE 2

Analysis of allele frequencies at several human MHC loci

Population	k	n	% nulls	$\hat{S}$	P_H
		HLA-A			
Danish	18	262	0	151.5	0.363
French	20	988	1.0	221.0	0.174
German	20	750	0.2	167.6	0.137
Italian	21	1106	2.5	175.5	0.031
Spanish	22	442	1.5	209.3	0.176
U.S. (Caucasian)	22	508	6.3	207.4	0.204
Canadian	19	434	0.7	160.4	0.182
Brazilian	22	674	3.6	236	0.058
Native American	19	320	4.3	83.6	0.919
Inuit	14	310	0.6	47.9	0.953
Japanese	18	1886	2.6	89.5	0.338
Korean	15	524	1.5	77.3	0.212
Han Chinese	18	358	9.3	132.9	0.376
Mongolian	18	268	4.2	222.8	0.049
Tibetan	11	198	3.9	29.5	0.377
Thai	22	482	4.0	163.7	0.395
Vietnamese	16	224	8.8	118.0	0.278
Senegalese	14	206	4.0	186.1	0.010
New Guinean	6	200	6.8	17.0	0.469
		HLA-B			
Danish	35		3.0	632.9	0.215
French	36		2.3	779.5	0.00028
German	38		0.0	687.7	0.005
Italian	39		2.2	175.5	0.031
Spanish	30		1.6	584.1	0.02
U.S. (Caucasian)	37		4.3	735.3	0.004
Canadian	34		2.0	588.0	0.067
Brazilian	36		2.7	767.6	0.003
Native American	23		1.7	218.2	0.124
Inuit	21		2.5	150.7	0.480
Japanese	34		2.1	425.4	0.001
Korean	28		1.5	544.1	0.00024
Han Chinese	31		5.4	517.1	0.117
Mongolian	32		2.9	1005.6	0.0001
Thai	30		0.0	467.3	0.005
Vietnamese	31		3.4	687.8	0.075
Senegalese	26		4.6	854.4	0.0009
New Guinean	17		2.9	106.8	0.572
Tibetan	29		7.0	423.8	0.204
		HLA-DRB1			
Danish	18		3.8	206.8	0.009
French	20		4.4	186.1	0.002
German	20		1.1	168.0	0.006
Italian	16		7.1	137.6	0.009
Spanish	17		3.5	135.2	0.023
U.S. (Caucasian)	21		5.0	194.9	0.007
Canadian	16		1.3	161.0	0.002
Brazilian	19		4.5	172.8	0.002
Native American	16		2.8	91.4	0.874
Inuit	15		9.0	105.6	0.301
Japanese	23		3.0	203.0	0.002
Korean	17		2.6	148.3	0.004
Han Chinese	16		6.2	115.4	0.024
Mongolian	15		8.0	282.7	0.0001
Thai	17		3.3	157.6	0.006
Vietnamese	16		2.7	179.6	0.159
		HLA-DQA1			
Danish	6		1.5	29.8	0.062
French	10		8.7	120.4	0.0003
German	10		1.4	168.0	0.006
Italian	10		9.9	46.0	0.021
Spanish	10		6.6	99.0	0.006
U.S. (Caucasian)	9		8.7	44.4	0.065
Canadian	9		3.3	32.6	0.113
Native American	10		6.9	42.1	0.834
Inuit	10		10.0	31.9	0.845
Japanese	11		10.0	36.3	0.030
Thai	10		4.2	49.5	0.127
Vietnamese	9		6.4	49.7	0.146

Data from Tsuji *et al.* (1992). The sample sizes, n, are the same for all loci and are not repeated in the latter parts of the table.

for which $\hat{S} < \infty$. For a particular data set, it appears that if an estimate can be obtained at all, it is of the correct order of magnitude at least.

In almost all cases, neutrality could be rejected using the Ewens-Watterson test even with a sample size of 100. For the examples summarized in Table 1 with $n = 100$, with $u = 10^{-6}$ and $s = 0.02$ ($S = 400$ and $M = 0.1$) $P_H < 0.05$ in 997 out of 1000 replicates and with $u = 10^{-7}$, $P_H < 0.05$ in all 1000 replicates. For the larger sample size ($n = 1000$) and both selection intensities ($S = 400$ and $S = 4000$), and for the smaller sample size ($n = 100$) and the stronger selection intensity ($S = 4000$), $P_H < 0.05$ for all 1000 replicates in each case.

We also simulated some cases with weaker selection, $S = 40$. For this selection coefficient, we had to use higher mutation rates to obtain numbers of alleles comparable to what is found in many data sets we analyzed. In all cases, $N = 10{,}000$. With $u = 10^{-5}$, the average number of alleles is ~9 and for $u = 5 \times 10^{-5}$, the average is ~15. For the smaller mutation rate, $P_H < 0.05$ in 603 of 1000 replicates with $n = 100$, and $P_H < 0.05$ in 785 of 1000 replicates with $n = 1000$. With the higher mutation rate, $P_H < 0.05$ in 318 of 1000 replicates with $n = 100$ and 453 of 1000 replicates with $n = 1000$. In all cases, averages of $\hat{S}$ were biased upward somewhat.

Even when the entire population is sampled, $\hat{S}$ is not the true value. The reason is that $\hat{S}$ reflects two sources of variation, variation attributable to sampling and variation attributable to stochastic variation in the allele frequencies. Sampling the entire population removes only one source of variation. The relatively weak dependence of the variance of $\hat{S}$ on sample size shows that stochastic variation is the more important.

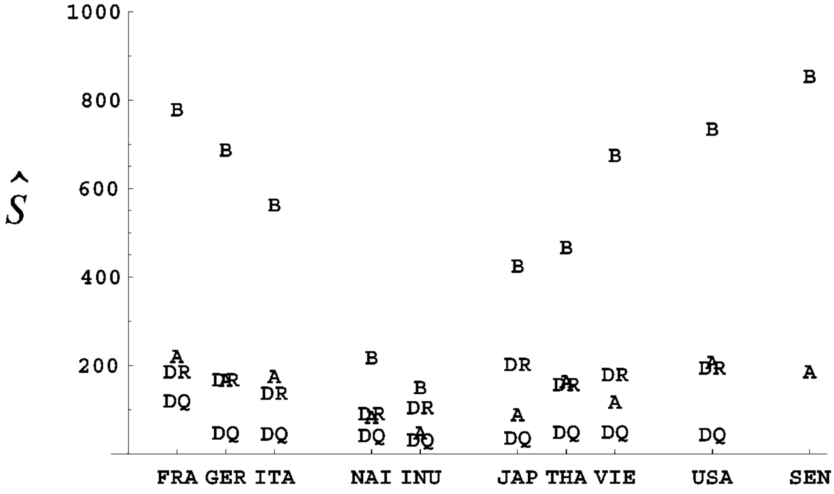

Figure 2.—Values of $\hat{S}$ for alleles distinguished by serotyping for four human *Mhc* loci (A, *HLA-A*; B, *HLA-B*; DQ, *HLA-DQA1*; DR, *HLA-DRB1*) in 11 populations (FRA, French; GER, German; ITA, Italian; NAI, Native American; INU, Inuit; JAP, Japanese; THA, Thai; VIE, Vietnamese; USA, United States; SEN, Senegalese).

Estimates of M obtained in the same set of replicates are not as good and are probably not very useful. In most of the cases we examined, average estimates of M are not of the correct order of magnitude and the 95% confidence interval covers several orders of magnitude. For example, with $n = 1000$, $S = 400$, and $M = 0.01$, the average estimate of M was 0.11 and the 95% confidence interval was (9.8×10^{-8}, 0.44). Our conclusion is that the method estimates S reasonably well, to within an order of magnitude at least, even for samples as small as 100, but that the estimate of M is not accurate enough to be useful.

APPLICATIONS

We applied our method for estimating S to two published data sets. The first data set is based on alleles distinguished by serological typing (or serotyping) and was presented at the histocompatibility workshop held in 1991 (Tsuji *et al.* 1992). These data were analyzed previously by Satta *et al.* (1994) and copies of these data were kindly provided to us by Dr. Y. Satta. We analyzed data from four loci, two class I loci, human histocompatibility system *HLA-A* and *HLA-B*, and two class II loci, *HLA-DQA1* and *HLA-DRB1*. The second data set was for class II loci for which alleles could be distinguished by differences in DNA sequence in the known or presumed peptide-binding regions (*molecular typing*), which are thought to be the targets of selection (Hughes and Yeager 1998). We analyzed the data for four loci, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPB1*, and *HLA-DRB1*. These data were published in the proceedings of a later workshop on HLA (Charron 1997) and were analyzed previously by Valdes *et al.* (1999). These data were kindly provided by Dr. S. McWeeney. Only alleles at class II loci can be distinguished by molecular typing.

Alleles distinguished by serotyping: We restricted our analysis to populations in which the sample size was >100 chromosomes and the proportion of null or unrecognizable alleles was usually <0.05. To obtain wider geographic coverage, we included some populations with slightly higher frequencies of null alleles. Table 2 shows the populations, loci, and sample sizes, and the values of $\hat{S}$ and P_{H} for each case. In no case was $\hat{S} = \infty$ obtained and in all cases P_{M} was 0 or nearly so, indicating that the hypothesis $S = \infty$ could always be rejected.

For *HLA-A*, values of P_{H} indicate neutrality cannot be rejected in most populations, yet values of $\hat{S}$ are generally larger than those from *HLA-DRB1* and *HLA-DQA1*, for which neutrality could usually be rejected. Neutrality could not be rejected for any of the samples from Native Americans and Inuits for any of the four loci. In those two populations, the value of P_{H} for *HLA-A* is close enough to one that there appears to be significantly more homozygosity than expected under neutrality. It is difficult to know whether the excess homozygosity results from a problem with serotyping, from evidence of purifying selection, or from admixture in those populations. We considered the possibility that a population bottleneck could lead to excess homozygosity, but simulation results from a model of a bottlenecked population could not reproduce the patterns found in the Native American and Inuit samples (results not shown). A bottleneck generally results in the loss of alleles rather than excess homozygosity.

Our results are similar to those obtained by Satta *et al.* (1994). Our estimates of S are consistently lower than those in Table 2 of Satta *et al.* (1994). Their estimates of S were based on the comparison of pairwise differences in silent and replacement sites in the peptide-binding region. They used five different methods that gave somewhat different results. Their method I gave the lowest estimates: *HLA-A*, $S = 690$; *HLA-B*, $S = 1200$; *HLA-DRB*, $S = 990$; and *HLA-DQB1*, $S = 360$. For *HLA-A*, the largest values of $\hat{S}$ we found are slightly >200 (Figure 2). For *HLA-B*, most of the $\hat{S}$ values were between 500 and 800 with only one >1000. The ratio of $\hat{S}$ for *HLA-B* to *HLA-A* is in the range 2.5–4, which is slightly higher

TABLE 3

Analysis of allele frequencies at four human MHC loci based on sequence typing of alleles

Population	k	n	$\hat{S}$	P_H	Population	k	n	$\hat{S}$	P_H
		HLA-DQA1					*HLA-DPB1*		
Belgian	6	80	56.4	0.044	Belgian	14	80	105.9	0.701
Bulgarian	8	84	47.1	0.198	Bulgarian	9	84	50.3	0.329
Croatian	8	186	49.8	0.018	Croatian	15	180	81.6	0.805
Czech	7	54	85.0	0.081	Czech	9	54	53.6	0.512
Greek Cypriot	9	190	42.9	0.183	Greek Cypriot	16	192	76.6	0.788
Gypsy	6	68	43.8	0.052	Gypsy	13	68	96.4	0.733
Italian	10	108	88.7	0.097	Italian	15	110	103.2	0.540
Northern French	7	466	26.1	0.026	Lebanese	22	342	132.7	0.795
Punjabi	8	102	40.9	0.057	Northern French	21	466	151.2	0.612
Spanish	10	138	134.2	0.003	Punjabi	19	102	152.1	0.957
Slovenian	10	186	63.4	0.103	Spanish	29	136	661.3	0.488
Guarani	7	178	15.8	0.549	Swiss	17	168	136.7	0.624
Kaingang	5	204	11.9	0.206	Slovenian	16	186	79.3	0.902
Mixe	4	106	11.5	0.221	Guarani	8	176	27.5	0.282
Mixtec	5	206	20.6	0.119	Kaingang	7	206	25.4	0.471
Zapotec	6	164	22.8	0.125	Mixe	5	108	8.4	0.960
Merina	7	326	36.9	0.051	Mixtec	10	206	29.6	0.964
Trobriand	5	150	15.4	0.079	Zapotec	12	182	42.7	0.980
Uygur	8	54	76.6	0.059	Merina	15	326	116.0	0.044
					Uygur	12	48	141.1	0.403
		HLA-DQB1					*HLA-DRB1*		
Belgian	12	80	201.3	0.020	Belgian	22	82	534.8	0.203
Bulgarian	11	84	96.4	0.124	Bulgarian	22	92	383.2	0.449
Croatian	13	180	128.7	0.036	Croatian	27	154	718.8	0.008
Czech	11	54	178.3	0.106	Czech	20	68	816.4	0.047
Greek Cypriot	12	182	87.0	0.038	Greek Cypriot	28	200	529.2	0.104
Gypsy	10	68	176.2	0.029	Italian	24	194	380.0	0.052
Italian	13	110	190.2	0.052	Lebanese	40	510	597.2	0.338
Lebanese	11	334	78.3	0.131	Northern French	37	468	523.7	0.102
Northern French	15	458	85.3	0.040	Swiss	25	142	450.9	0.180
Punjabi	12	74	287.7	0.010	Slovenian	24	200	314.1	0.036
Spanish	11	126	135.5	0.008	Kaingang	4	56	13.0	0.215
Slovenian	15	186	132.9	0.033	Mixe	9	96	38.5	0.583
Guarani	7	150	19.6	0.557	Mixtec	14	198	73.7	0.388
Kaingang	5	172	12.1	0.244	Merina	23	286	292.8	0.459
Mixe	5	108	13.2	0.138	Trobriand	14	156	83.0	0.463
Mixtec	10	206	34.2	0.390	Uygur	26	128	589.6	0.174
Zapotec	11	156	45.7	0.405	Zaire	23	180	381.6	0.064
Merina	11	312	46.7	0.283					
Uygur	13	54	182.4	0.599					

Data are from Charron (1997). The order of the populations is the same as in Table 3 of Valdes *et al.* (1999).

than the ratio found by Satta *et al.* for their method I but comparable to values they obtained using the other four methods. Given the wide confidence intervals associated with both methods for estimating selection intensities, our results are compatible with those of Satta *et al.* (1994).

Alleles distinguished by molecular typing: Table 3 shows the results for four class II loci. Values of $\hat{S}$ are consistently larger for *DQB1* than *DQA1*. Values for *DRB1* are the largest. The orders of magnitude of the values of $\hat{S}$ for the loci are the same as found by Satta *et al.* (1994) for the same four loci, but, as with the previous data set, our estimates of S are consistently smaller than theirs. For *DQA1* and *DRB1* we have estimates based on both serotyping and molecular typing. The data sets are mostly for different populations and, even when the same population is represented, different individuals were sampled. The results from analyzing the two data sets are comparable. Most values of $\hat{S}$ for European populations are between 50 and 100 with smaller values for Native American populations. The values of $\hat{S}$ based on molecular typing are somewhat

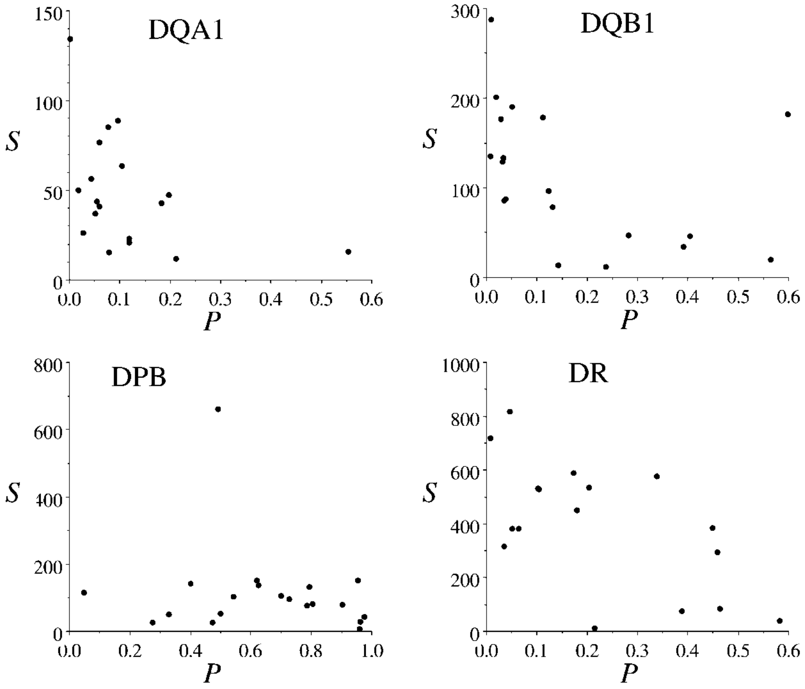

Figure 3.—Values of $\hat{S}$ and P_H from Table 3 for alleles distinguished by molecular typing for four *Mhc* class II loci in several human populations.

smaller than based on serotyping for Native Americans. Values of $\hat{S}$ are larger for *DRB1*, and again Native American populations have the smallest values.

DISCUSSION AND CONCLUSIONS

Our method provides a simple way to estimate the strength of overdominant selection from the observed distribution of allele frequencies. The estimate of S is essentially the inverse of the variance in frequencies. The multinomial Dirichlet distribution (3) is used to account for sampling. Our estimate of S is always an underestimate of the true intensity of selection because other factors, particularly differences in fitness among different alleles or anything else that causes deviations from the stationary distribution of allele frequencies, increase the variance and hence reduce the estimate of S. Our method assumes all variation in observed allele frequencies is the result of sampling and genetic drift.

The estimate of S obtained is an estimate of the net force tending to equalize allele frequencies. That force could be attributable only to overdominance in fitness, as we have assumed, or to frequency-dependent selection resulting from mating preferences or some other factor. The estimate itself does not indicate the kind of selection acting and does not validate a particular model of selection.

Another possible cause of differences in allele frequencies is bias in the mutation process, with some alleles or types of alleles arising more frequently than others. Intragenic recombination between different alleles could contribute to bias in mutation because recombination between alleles that are more different in DNA sequence in the peptide-binding region would be more likely to create functionally different alleles. To explore the effects of intragenic recombination, a sequence-based simulation model of the kind analyzed by Satta (1997) is necessary but that is beyond the scope of our study. Satta assumed that heterozygote fitness increased with the difference in DNA sequence of the peptide-binding region and allowed for intragenic recombination to create functionally different alleles. Both of these assumptions lead to increased variance in allele frequencies under selection alone and hence would reduce estimates of S obtained using our method.

Our method is similar to methods developed by M. N. Grote (unpublished results) and by P. Donnelly, M. Nordberg and P. Joyce (unpublished results), both of which rely on the theory of Joyce (1995). These methods are mathematically much more sophisticated than the method presented here and require extensive simulation to obtain an estimate of S for a single data set. Donnelly *et al.* apply their method to data for a locus with four alleles and observed counts (46, 166, 112, and 120) and estimate S to be 36. Our method gives 47, which is consistent with their results.

There is no simple relationship between our estimates of S and the tail probability from the Ewens-Watterson homozygosity test, P_H. The variance in allele frequency is closely related to the homozygosity,

$$\mathrm{Var}(x) = \frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2 = \frac{1}{k}\left(F - \frac{1}{k}\right), \qquad (6)$$

where k is the number of alleles, x_i is the frequency of the ith allele, $\overline{x} = 1/k$ is the average allele frequency, and $F = \Sigma_{i=1}^{k} x_i^2$ is the homozygosity. But because of sampling, $\hat{S}$ and P_H depend in complex ways on k and the counts of each allele. Figure 3 shows plots of $\hat{S}$ *vs.* P_H for each of the four loci in Table 3. Although it would be convenient if P_H indicated roughly the intensity of selection, it does not appear to do so very well.

Our estimates of S for *HLA-A* are of the same order of magnitude as those obtained by Satta *et al.* (1994) from the analysis of silent and replacement differences in the peptide-binding region of *HLA-A*. There appears to be some geographic variation in the intensities of overdominant selection in human populations. The broad confidence intervals on $\hat{S}$ we found in simulations mean that differences among European, Asian, and the one African population analyzed are not significant at any of the loci. It does appear that in the two North American populations, Native Americans and Inuit, either selection is much weaker than in the rest of the world or that admixture has resulted in the appearance of weaker selection.

We thank Dr. Y. Satta for providing copies of the data analyzed in the article by Satta *et al.* (1994) and Dr. S. McWeeney for providing copies of the data analyzed by Valdes *et al.* (1999), and for helpful discussions of this project. This research was supported in part by National Institutes of Health (NIH) grant GM40282 to M. Slatkin and by an NIH traineeship to C. A. Muirhead from grant GM07127.

LITERATURE CITED

Charron, D. (Editor), 1997 *Genetic Diversity of HLA: Functional and Medical Implications.* EDK Press, Paris.

Edwards, S. V., and P. W. Hedrick, 1998 Evolution and ecology of MHC molecules: from genomics to sexual selection. Trends Ecol. Evol. **13:** 305–311.

Felsenstein, J., 1977 Multivariate normal genetic models with a finite number of loci, pp. 227–246 in *Proceedings of the International Conference on Population Genetics*, edited by E. Pollak, O. Kempthorne and T. J. Bailey. Iowa State University Press, Ames, IA.

Hughes, A. L., and M. Yeager, 1998 Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet. **32:** 415–435.

Johnson, N. L., S. Kotz and N. Balakrishnan, 1997 *Discrete Multivariate Distributions.* John Wiley & Sons, New York.

Joyce, P., 1995 Robustness of the Ewens sampling formula. J. Appl. Probab. **32:** 609–622.

Satta, Y., 1997 Effects of intra-locus recombination of HLA polymorphism. Hereditas **127:** 105–112.

Satta, Y., C. O'Huigin, N. Takahata and J. Klein, 1994 Intensity of natural selection at the major histocompatibility complex loci. Proc. Natl. Acad. Sci. USA **91:** 7184–7188.

Slatkin, M., 1994 An exact test for neutrality based on the Ewens sampling distribution. Genet. Res. **64:** 71–74.

Slatkin, M., 1996 A correction to the exact test based on the Ewens sampling distribution. Genet. Res. **68:** 259–260.

Slatkin, M., and C. A. Muirhead, 1999 Overdominant alleles in a population of variable size. Genetics **152:** 775–781.

Takahata, N., 1990 A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc. Natl. Acad. Sci. USA **87:** 2419–2423.

Tsuji, K., M. Aizawa and T. Sasazuki (Editors), 1992 *HLA 1991: Proceedings of the Eleventh International Histocompatibility Workshop and Conference Held in Yokohama, Japan, 6-13 November, 1991.* Oxford University Press, Oxford.

Valdes, A. M., S. K. McWeeney, D. Meyer, M. P. Nelson and G. Thomson, 1999 Locus and population specific evolution in HLA class II genes. Ann. Hum. Genet. **63:** 27–43.

Watterson, G. A., 1977 Heterosis or neutrality? Genetics **85:** 789–814.

Wolfram, S., 1996 *The Mathematica Book.* Wolfram Media, Champaign, IL.

Communicating editor: N. Takahata