THE SAMPLING DISTRIBUTION OF LINKAGE DISEQUILIBRIUM UNDER AN INFINITE ALLELE MODEL WITHOUT SELECTION

RICHARD R. HUDSON

Biometry and Risk Assessment Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

Manuscript received June 25, 1984
Revised copy accepted November 16, 1984

ABSTRACT

The sampling distributions of several statistics that measure the association of alleles on gametes (linkage disequilibrium) are estimated under a two-locus neutral infinite allele model using an efficient Monte Carlo method. An often used approximation for the mean squared linkage disequilibrium is shown to be inaccurate unless the proper statistical conditioning is used. The joint distribution of linkage disequilibrium and the allele frequencies in the sample is studied. This estimated joint distribution is sufficient for obtaining an approximate maximum likelihood estimate of $C = 4Nc$, where N is the population size and c is the recombination rate. It has been suggested that observations of high linkage disequilibrium might be a good basis for rejecting a neutral model in favor of a model in which natural selection maintains genetic variation. It is found that a single sample of chromosomes, examined at two loci cannot provide sufficient information for such a test if $C < 10$, because with C this small, very high levels of linkage disequilibrium are not unexpected under the neutral model. In samples of size 50, it is found that, even when C is as large as 50, the distribution of linkage disequilibrium conditional on the allele frequencies is substantially different from the distribution when there is no linkage between the loci. When conditioned on the number of alleles at each locus in the sample, all of the sample statistics examined are nearly independent of $\theta = 4N\mu$, where μ is the neutral mutation rate.

IT has been suggested that the correlation of alleles on gametes (linkage disequilibrium) may be a sensitive indicator of the action of natural selection (LEWONTIN 1964, 1974). It is known from the analysis of multilocus models that selection can produce strong correlations among alleles at different loci, even without strong epistatic interactions (FRANKLIN and LEWONTIN 1970). Unfortunately, the distribution of linkage disequilibrium has not been well characterized under models without selection. This has made it difficult to interpret observations from natural populations. Both hypothesis testing and estimation require better knowledge of the sampling distribution of linkage disequilibrium than is currently available. To learn more about the distribution of linkage disequilibrium a series of Monte Carlo simulations was carried out using a neutral model which is relevant for interpreting observations on natural populations. The results of those simulations are reported here.

STATISTICAL QUANTITIES TO BE STUDIED

The model considered is a two-locus Wright-Fisher infinite allele model with random union of gametes (KARLIN and MCGREGOR 1968). In this model generations are discrete, population size is constant and multinomial sampling with mutation and recombination produce succeeding generations. It is assumed that no population subdivision occurs. Some properties of linkage disequilibrium are known for this model and will be briefly reviewed here. See EWENS (1979) for a detailed description of the model and a more complete review of earlier work. Under this model many alleles may be present at each of the loci. Label the alleles at one locus A_1, A_2 . . . and the alleles at the other locus B_1, B_2 . . . . The most common measure of the association of A_i with B_j on gametes is:

$$D_{ij} = f_{ij} - p_i q_j, \tag{1}$$

where f_{ij} is the population frequency of A_iB_j gametes, p_i is the frequency of the A_i allele and q_j is the frequency of the B_j allele. The expected value of the statistic

$$D^2 = \sum\sum D_{ij}^2 \tag{2}$$

is known for the model just described and the related k allele model (HILL 1976; GRIFFITHS 1981; TAKAHATA 1982). $E(D^2)$ is a function of $\theta = 4N\mu$ and $C = 4Nc$, where N is the diploid population size, μ is the neutral mutation rate at each locus and c is the recombination rate between the two loci. Recently, GOLDING and STROBECK (1983), using a numerical method, found that the variance of D^2 can be quite large. This led them to doubt the usefulness of D^2 as a test statistic. Other statistics that have been used to measure the association of alleles on gametes are:

$$r_{ij} = D_{ij}/[p_i q_j (1 - p_i)(1 - q_j)]^{1/2}, \tag{3}$$

$$r^2 = D^2/[(1 - F_A)(1 - F_B)], \tag{4}$$

where

$$F_A = \sum p_i^2 \quad \text{and} \quad F_B = \sum q_j^2$$

and

$$D'_{ij} = D_{ij}/D_m, \tag{5}$$

where

$$D_m = \begin{cases} \min[p_i q_j, (1 - p_i)(1 - q_j)] \text{ if } D_{ij} < 0 \\ \min[(1 - p_i)q_j, p_i(1 - q_j)] \text{ otherwise.} \end{cases}$$

The expectation of r^2 is not known, but it has been suggested that the standard linkage disequilibrium squared, $\sigma_d^2 = E(D^2)/E((1 - F_A)(1 - F_B))$, would closely approximate the expectation of r^2 (OHTA and KIMURA 1969, 1971; KIMURA and OHTA 1971; HILL 1975). For the model being considered, σ_d^2 can be calculated using equation (10) by HILL (1975). MARUYAMA (1982) used

simulations to estimate $E(r^2)$ and showed that σ_d^2 can differ substantially from $E(r^2)$. The statistic D'_{ij} was introduced by LEWONTIN (1964). Unlike D_{ij} and r_{ij}, the range of values that D'_{ij} can take is not dependent on p_i or q_j. When there are two alleles at each locus and no recombination takes place, D'_{ij} is always 1 or -1.

The statistics D^2, D_{ij}, r_{ij} and D'_{ij} are defined in terms of population frequencies of alleles and gametes and, thus, are not directly observable in practice. For hypothesis testing and estimation, it is the distribution of sample statistics that must be characterized. Sample statistics analogous to the population statistics defined in (1)–(5) are the focus of this study. The following notation will be used. Sample statistics are designated with a tilde ($\sim$) above the symbol. The number of alleles in the sample at the A locus is denoted $\tilde{k}_A$; the number of alleles at the B locus is $\tilde{k}_B$. It is convenient to number the alleles in order of decreasing frequency in the sample, so that $\tilde{p}_1 \geq \tilde{p}_2 \geq \ldots$ and $\tilde{q}_1 \geq \tilde{q}_2 \geq \ldots$. With this labeling, $\tilde{D}_{11}$ is the sample disequilibrium between the two most frequent alleles in the sample. The sample size is denoted n. Several joint and conditional distributions are considered. To simplify notation, the following convention is adopted. For any collection of random variables, $x, y, \ldots$ and events $A, B, \ldots$, the joint distribution of the random variables $x, y, \ldots$ and the events $A, B, \ldots$ is referred to as the distribution of $(x, y, \ldots; A, B, \ldots)$. The distribution of the random variables $x, y, \ldots$ conditional on the events $A, B, \ldots$ is referred to as the distribution of $(x, y, \ldots \mid A, B, \ldots)$. For example, the distribution of $(\tilde{r}^2; \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$ refers to the joint distribution of $\tilde{r}^2$ and the events $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 \leq 0.95$ and $\tilde{q}_1 \leq 0.95$. The distribution of $\tilde{r}^2$ conditional on $\tilde{k}_A = \tilde{k}_B = 2$ is referred to as the distribution of $\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2$. The expectation of a random variable, x is denoted $E(x)$. The expectation of x conditional on $A, B, \ldots$ is denoted $E(x \mid A, B, \ldots)$. Estimates of these expectations, obtained from Monte Carlo simulations are denoted $\bar{x}$ and $\bar{x} \mid A, B, \ldots$, respectively.

Even less is known about the sample statistics $\tilde{r}^2$, $\tilde{r}_{11}$, $\tilde{D}^2$, $\tilde{D}_{11}$ and $\tilde{D}'_{11}$ than about the analogous population statistics. Recently, however, GOLDING (1984) has used a numerical method to obtain the distribution of $(\tilde{D}_{11}; \tilde{k}_A = \tilde{k}_B = 2)$ and $(\tilde{r}_{11}; \tilde{k}_A = \tilde{k}_B = 2)$ for $C = 0$ and $C = \infty$.

The primary objective of this study is the estimation of the distributions of the sample statistics $\tilde{r}^2$, $\tilde{D}^2$, $\tilde{r}_{11}$, $\tilde{D}_{11}$ and $\tilde{D}'_{11}$ for several levels of recombination. Because many loci in natural populations exhibit only two alleles in samples, most of the distributions considered are conditional on, or jointly with, $\tilde{k}_A = \tilde{k}_B = 2$. Also, in some studies of linkage disequilibrium in natural populations (*e.g.*, LANGLEY, TOBARI and KOJIMA 1974), monomorphic loci or nearly monomorphic loci are not considered. This is reasonable since monomorphic loci, or nearly monomorphic loci, provide little or no information concerning the amount of association between alleles at different loci. But to interpret the observations made in such studies requires that one consider the distributions of sample statistics conditional on minimal levels of polymorphism in the sample. This motivated the estimation of the distributions of statistics conditional on, or jointly with, $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 \leq 0.95$, and $\tilde{q}_1 \leq 0.95$. Also, to investigate

the possibility that more information could be obtained from a sample by considering the joint distribution of linkage disequilibrium and the allele frequencies in the sample, the distribution of $(\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1; \tilde{k}_A = \tilde{k}_B = 2)$ was estimated for several levels of recombination and several combinations of values of $\tilde{p}_1$ and $\tilde{q}_1$.

A formula for the expectation of the sample statistic $\tilde{D}^2$ is derived in the APPENDIX. The expectation of $\tilde{D}^2$ and $E(D^2)$ are compared for a range of parameter values. For several parameter combinations the following expectations were estimated with Monte Carlo simulations: $E(\tilde{D}^2)$, $E(\tilde{D}^2 \mid \tilde{k}_A = \tilde{k}_B = 2)$, $E(\tilde{D}^2 \mid k_A = k_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$, $E(\tilde{r}^2)$, $E(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2)$ and $E(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$. In addition, estimates were obtained of the distributions of the following:

(a) $\tilde{r}^2$,
(b) $(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$,
(c) $(\tilde{r}_{11}; \tilde{k}_A = \tilde{k}_B = 2)$
(d) $(\tilde{r}_{11}; \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.9, \tilde{q}_1 \leq 0.9)$,
(e) $(\tilde{D}_{11}; \tilde{k}_A = \tilde{k}_B = 2)$,
(f) $(\tilde{D}_{11}; \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.9, \tilde{q}_1 \leq 0.9)$,
(g) $(\tilde{D}'_{11}; \tilde{k}_A = \tilde{k}_B = 2)$,
(h) $(\tilde{D}'_{11}; \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.9, \tilde{q}_1 \leq 0.9)$,
(i) $(\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1; \tilde{k}_A = \tilde{k}_B = 2)$.

The estimated distributions are used to address several questions. How well is $E(\tilde{r}^2)$ approximated by σ_d^2 or even $E(r^2)$? How similar are the distributions of r^2, $\tilde{r}^2$ and $(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$? Similarly, how different are the distributions of $(\tilde{r}_{11} \mid \tilde{k}_A = \tilde{k}_B = 2)$ and $(\tilde{r}_{11} \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.9, \tilde{q}_1 \leq 0.9)$? The expectation of D^2 is strongly dependent on θ. How sensitive to θ are the sample statistics when conditioned on $\tilde{k}_A = \tilde{k}_B = 2$? How precise an estimate of C can be made with a single sample of gametes that have been examined at just two loci? Are measures of linkage disequilibrium likely to be useful as test statistics? Which statistics are most informative?

The algorithm used to produce samples under the neutral model is briefly described in the next section. Estimates of the sample distributions are presented and described in RESULTS. The implications of distributions for hypothesis testing and estimation are presented in DISCUSSION.

SIMULATION METHODS

The distributions of the sample statistics were estimated with a Monte Carlo method which is quite different from the standard method that requires that the entire population be represented in computer memory. With the method used in this study, a random sample of gametes is generated in a two-phase process. In the first phase, the history of the sample of gametes is generated. For a two-locus model, the history can be represented by two trees, one tree for each locus. Such a tree specifies which gametes are most closely related and at what time the most recent common ancestor of any pair of the gametes

occurred. For two linked loci the two trees are obviously correlated. The generation of these correlated trees is described in detail by HUDSON (1983). In the second phase, the numbers of mutations that occur on each branch of the tree are generated. Given the mutation rate and the duration of the branches, the numbers of mutations are generated, assuming that the numbers of mutations have a Poisson distribution. With the trees and the numbers of mutations on each branch, the allelic composition of each gamete is determined and any sample statistic can be calculated. This method of generating samples relies on the assumption that the population size (N) is large and recombination rate (c) and mutation rate (μ) are small.

To study the distribution of some statistics a slightly different technique is used. The history of a sample of gametes is generated as before, but then, instead of generating a single sample from the given pair of trees, the entire distribution of the statistic is calculated conditional on that particular pair of trees. The distribution of the statistic is then estimated by averaging such conditional distributions from many such pairs of trees. To illustrate the technique, consider the following method of estimating the distribution of $(\tilde{D}_{11};\ \tilde{k}_A = \tilde{k}_B = 2,\ \tilde{p}_1 = p,\ \tilde{q}_1 = q)$. To estimate this distribution many pairs of trees were generated as described before. For each pair of trees the distribution of $(\tilde{D}_{11};\ \tilde{k}_A = \tilde{k}_B = 2,\ \tilde{p}_1 = p,\ \tilde{q}_1 = q)$ conditional on that pair of trees was determined as follows. Notice that, if two alleles are segregating in the sample at locus A, it must be the case that one branch of the A locus tree has one or more mutations on it, and all of the other branches have none. The branches of the tree were examined to determine which ones, if any, were such that mutations occurring on them would result in alleles in the sample at frequency p and $1 - p$. Denote the set of such branches B_p. The probability that one or more mutations occur on a particular branch and none occur elsewhere on the tree is just $(1 - e^{-\mu t})e^{-\mu(T-t)}$, where t is the duration of the particular branch and T is the sum of the durations of all of the branches. Similarly, the B locus tree was examined to find those branches such that mutations occurring on them would result in alleles of frequency q and $1 - q$. Denote this set of branches B_q. A sample would have two alleles at each locus with the specified frequencies if one or more mutations occur on one of the branches of set B_p, none occur elsewhere on the A locus tree, one or more mutations occur on one of the branches of set B_q and no mutations occur elsewhere on the B locus tree. For each way that this can occur, the probability was calculated and the value of $\tilde{D}_{11}$ that would result from the mutations was ascertained. In this way the distribution of $\tilde{D}_{11}$ was determined for a particular pair of trees. The average of such distributions over many pairs of trees is an estimate of the distribution of $(\tilde{D}_{11};\ \tilde{k}_A = \tilde{k}_B = 2,\ \tilde{p}_1 = p,\ \tilde{q}_1 = q)$.

The computer programs were checked against several known properties of the two-locus model. Specifically, the programs correctly generated the expected two-locus homozygosity (STROBECK and MORGAN 1978), the expectation of $\tilde{D}^2$ (see Table 1), and also the distributions of $\tilde{D}_{11}$ and $\tilde{r}_{11}$ for $C = 0$ (GOLDING 1984).

Listings of programs to produce samples of gametes under the two-locus

TABLE 1

Expected linkage disquilibrium in populations, in samples and in samples conditional on segregation

		Theoretical[a]		Simulation means				
					$\bar{\tilde{D}}^2 \mid \tilde{k}_A = \tilde{k}_B = \tilde{2})$		$\bar{\tilde{D}}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95$	
θ	C	$E(D^2)$	$E(\tilde{D}^2)$ $n = 50$	$\bar{\tilde{D}}^2$ $n = 50$	$n = 50$	$n = 100$	$n = 50$	$n = 100$
0.02	0	0.00021	0.00021	0.00019	0.025		0.054	0.050
	2	0.000095	0.000099	0.0001	0.012		0.028	0.025
	10	0.000031	0.000037	0.000033	0.0048		0.011	0.0090
	20	0.000017	0.000024	0.000024	0.0030		0.0068	0.0051
0.1	0	0.0040	0.0040	0.0038	0.024	0.019	0.055	0.050
	2	0.0019	0.0020	0.0020	0.013	0.0093	0.028	0.025
	10	0.00065	0.00079	0.00076	0.0048	0.0036	0.010	0.0091
	20	0.00036	0.00051	0.00049	0.0030	0.0021	0.0065	0.0051
0.2	0	0.012	0.012	0.012	0.025	0.018	0.055	0.050
	2	0.0062	0.0064	0.0063	0.013	0.0095	0.028	0.026
	10	0.0022	0.0026	0.0025	0.0049	0.0034	0.011	0.0091
	20	0.0012	0.0017	0.0017	0.0033	0.0018	0.0066	0.0052
0.4	0	0.029	0.029	0.029	0.023	0.017	0.054	0.051
	2	0.016	0.017	0.017	0.012	0.0094	0.03	0.026
	10	0.0061	0.0074	0.0073	0.0049	0.0035	0.011	0.0093
	20	0.035	0.0049	0.0049	0.0030	0.0021	0.0069	0.0053

[a] $E(D^2)$ calculated with equation (10) from HILL (1975); $E(\tilde{D}^2)$ calculated with (A7)–(A10) from the APPENDIX.

neutral model are available on request. The programs are written in the programming language C.

RESULTS

Table 1 shows $E(D^2)$, $E(\tilde{D}^2)$ and $\bar{\tilde{D}}^2$ for several combinations of θ, C and sample size. $E(D^2)$ is calculated using the formula of HILL (1975). $E(\tilde{D}^2)$ is calculated using a formula derived in the APPENDIX. For these parameter values, $E(D^2)$ differs only slightly from $E(\tilde{D}^2)$. $\bar{\tilde{D}}^2$ is the mean value of $\tilde{D}^2$ in a large number of computer-generated samples. $\tilde{D}^2$ is shown only to indicate that the computer program works correctly. Also shown are estimates of two conditional expectations, $E(\tilde{D}^2 \mid \tilde{k}_A = \tilde{k}_B = 2)$ and $E(\tilde{D}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$. Table 1 shows that conditioning on minimum levels of polymorphism can increase the mean value of $\tilde{D}^2$ greatly. Also note that the conditional expectations of $\tilde{D}^2$ are quite insensitive to θ.

Table 2 shows σ_d^2, $\bar{\tilde{r}}^2$ and $(\bar{\tilde{r}}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$. Also shown are estimates of $E(r^2)$ taken from table 2 of MARUYAMA (1982). As Maruyama pointed out σ_d^2 is not always a good predictor of $E(r^2)$. The results in Table 2 demonstrate that $E(\tilde{r}^2)$ in not well predicted by either $E(r^2)$ or σ_d^2. For ex-

TABLE 2

Comparison of σ_d^2 and estimated expectations of $\tilde{r}^2$ in populations, in samples and in samples conditional on certain levels of polymorphism

				$\bar{\tilde{r}}^2$			$\bar{\tilde{r}}^2 \mid \hat{k}_A = \hat{k}_B = 2$			$\bar{\tilde{r}}^2 \mid \hat{k}_A = \hat{k}_B = 2, \hat{p}_1 \leq 0.95, \hat{q}_1 \leq 0.95$		
θ	C	σ_d^2	$\bar{r}^2$	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
0.02	0	0.445					0.211			0.40	0.38	0.38
	2	0.228					0.118			0.22	0.20	0.20
	10	0.079					0.059			0.096	0.084	0.082
	20	0.044					0.043			0.065	0.052	0.050
0.1	0	0.411	0.061	0.215	0.180	0.140	0.210	0.166	0.116	0.40	0.38	0.38
	2	0.220	0.028	0.124	0.101	0.082	0.121	0.091	0.068	0.22	0.20	0.21
	10	0.078		0.063	0.049	0.041	0.059	0.046	0.043	0.095	0.084	0.082
	20	0.044		0.044	0.033	0.027	0.043	0.030	0.026	0.064	0.052	0.050
0.2	0	0.376	0.110	0.225	0.188	0.160	0.216	0.162	0.118	0.40	0.38	0.39
	2	0.209	0.05	0.130	0.110	0.089	0.123	0.094	0.067	0.22	0.20	0.21
	10	0.077		0.063	0.051	0.043	0.058	0.049	0.035	0.096	0.085	0.084
	20	0.043		0.048	0.034	0.029	0.047	0.031	0.028	0.064	0.053	0.049
0.4	0	0.322	0.162	0.224	0.199	0.179	0.208	0.162	0.177	0.41	0.38	0.38
	2	0.192	0.091	0.139	0.121	0.109	0.114	0.097	0.066	0.23	0.21	0.21
	10	0.074		0.068	0.057	0.049	0.061	0.046	0.032	0.099	0.086	0.083
	20	0.042		0.050	0.038	0.032	0.044	0.033	0.019	0.065	0.054	0.050

σ_d^2 calculated with equation (10) from HILL (1975). $\bar{r}^2$ is from table 2 of MARUYAMA (1982).

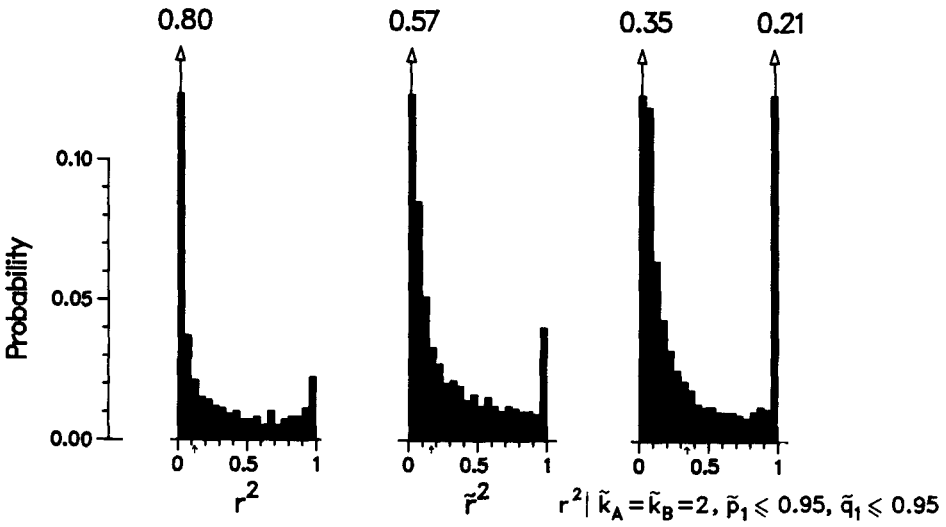

FIGURE 1.—A comparison of the distribution of the population statistic r^2 and the distributions of the two sample statistics $\tilde{r}^2$ and $\tilde{r}^2$ conditional on $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 < 0.95$ and $\tilde{q}_1 < 0.95$. For all three distributions $\theta = 0.1$ and $C = 0.2$. The sample statistics are for samples of size 100. The distribution of r^2 is taken directly from figure 5 of MARUYAMA (1982). The mean values of the statistics are shown by the small arrows along the horizontal axis.

ample, when $C = 2$ and $\theta = 0.1$, then $\sigma_d^2 = 0.22$, $\bar{r}^2 = 0.028$, and for $n = 100$, $\bar{\tilde{r}}^2 = 0.10$. Table 2 shows that the expectation of $\tilde{r}^2$ is substantially increased when conditioned on $\tilde{p}_1 \leq 0.95$ and $\tilde{q}_1 \leq 0.95$. For $C = 2$, and $n = 100$, the mean of $(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$ is approximately 0.20. The conditional expectation of $\tilde{r}^2$ appears to be quite insensitive to θ.

The distributions of r^2, $\tilde{r}^2$ *and* $(\tilde{r} \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$ are compared in Figure 1 for $\theta = 0.1$ and $C = 0.2$. The distribution of r^2 is from MARUYAMA (1982). The estimated mean values of these statistics are 0.10, 0.17 and 0.35, respectively. The three statistics differ considerably in their probabilities of taking values near 0 and 1. The estimated probabilities of taking a value less than 0.05 are 0.8, 0.57 and 0.35 for r^2, $\tilde{r}^2$ and $(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$, respectively. The probabilities of a value greater than 0.95 are 0.02, 0.04 and 0.21 for r^2, $\tilde{r}^2$ and $(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$, respectively.

Figures 2–4 show the distributions of $(\tilde{D}_{11}; \tilde{k}_A = \tilde{k}_B = 2)$, $(\tilde{r}_{11}; \tilde{k}_A = \tilde{k}_B = 2)$, and $(\tilde{D}'_{11}; \tilde{k}_A = \tilde{k}_B = 2)$, respectively, for samples of size 50 and $\theta = 0.1$. For $C = 0$, GOLDING (1984) obtained the distributions of $\tilde{D}_{11}$ and $\tilde{r}_{11}$ for these same parameter values. (Golding actually considers the linkage disequilibrium between the alleles present on the most frequent type of gamete rather than between the two most frequent alleles. The resulting distributions are only slightly different.) Also shown in Figures 2–4 are the distributions with $\tilde{p}_1 \leq 0.9$ and $\tilde{q}_1 \leq 0.9$.

Figure 2 shows that, for all levels of recombination, the distribution of $\tilde{D}_{11}$ has a large peak at $\tilde{D}_{11} \approx -0.01$. For $C = 0$, small positive values of $\tilde{D}_{11}$ are

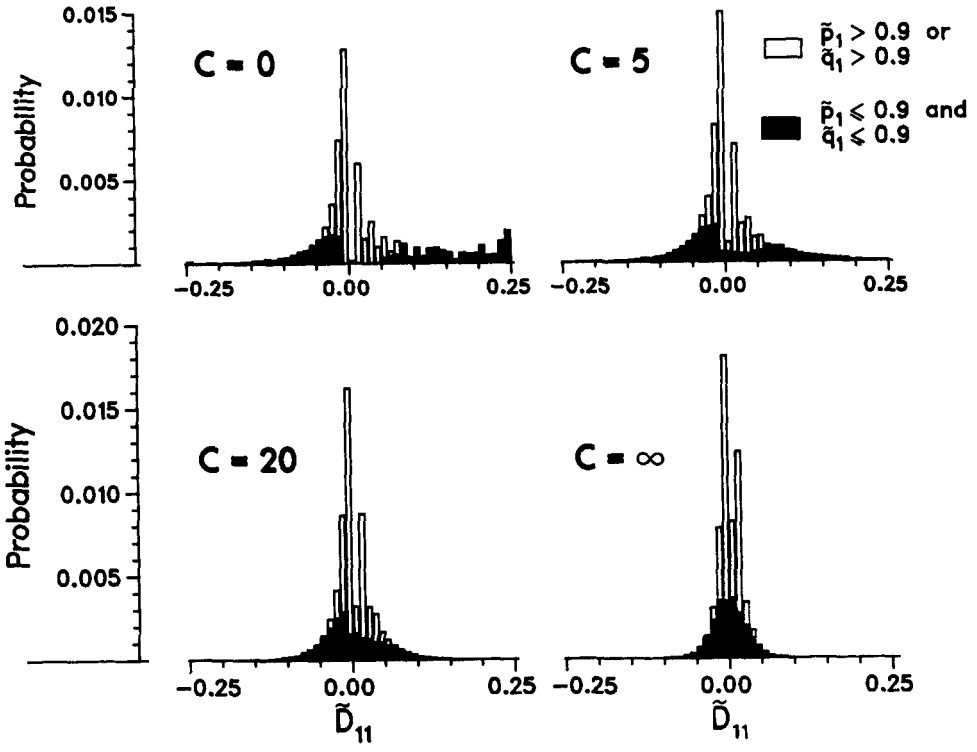

FIGURE 2.—Estimates of the distribution of $\tilde{D}_{11}$ at four levels of recombination ($\theta = 0.1$, $n = 50$). The total height of each column indicates the probability that $\tilde{k}_A = \tilde{k}_B = 2$ and $\tilde{D}_{11}$ takes a value in the interval covered by the base of the column. The heights of the solid black columns indicate the estimated probability that $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 \leq 0.9$, $\tilde{q}_1 \leq 0.9$ and $\tilde{D}_{11}$ belongs to the indicated interval. For $\theta = 0.1$ and $n = 50$, the probabilities of $\tilde{k}_A = \tilde{k}_B = 2$ were estimated to be 0.085, 0.83, 0.084 and 0.083 for $C = 0$, $C = 5$, $C = 20$ and $C = \infty$, respectively. The probabilities of ($\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 \leq 0.9$ and $\tilde{q}_1 \leq 0.9$) are approximately 0.026, 0.023, 0.023 and 0.022 for $C = 0$, 5, 20 and ∞, respectively.

very unlikely, but there is a substantial tail on the distribution that extends out to $\tilde{D}_{11} = 0.25$. As the recombination rate increases, the long tail disappears and the small gap in the distribution just to the right of 0 gradually fills in. If only samples with $\tilde{p}_1 \leq 0.9$ and $\tilde{q}_1 \leq 0.9$ are considered, the distributions are quite different. For $C = 0$, the large peak just to the left of 0 is almost completely eliminated, leaving a much more broadly distributed statistic with the tail to the right being a much more significant part of the distribution.

Figure 3 shows that the distribution of $\tilde{r}_{11}$, like the distribution of $\tilde{D}_{11}$, shows a large peak just to the left of 0 for all levels of recombination. For $C = 0$, $\tilde{r}_{11}$ is very likely to be either 1 or a small negative number. Small positive values are very unlikely. As recombination rates increase, the large probability mass at 1 disappears and the gap to the right of 0 fills in. Again, the distributions among samples with $\tilde{p}_1 \leq 0.9$ and $\tilde{q}_1 \leq 0.9$ are quite different. For such samples, with $C = 0$, there is no large peak just to the left of 0, whereas the large probability mass at $\tilde{r}_{11} = 1$ remains.

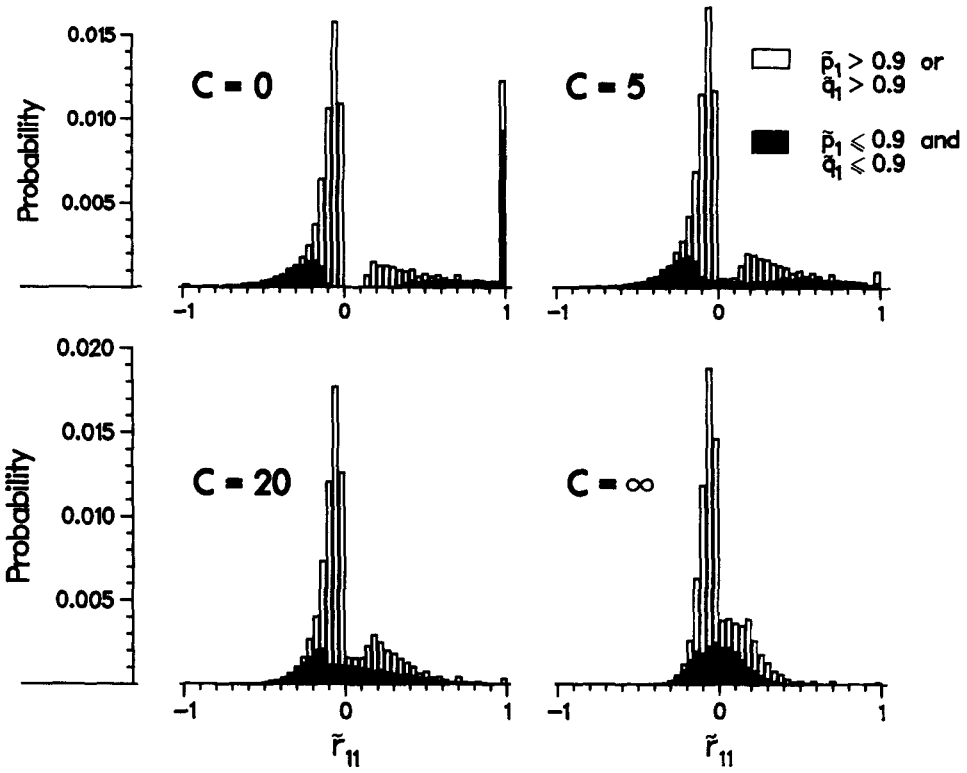

FIGURE 3.—Estimates of the distribution of $\tilde{r}_{11}$ at four levels of recombination ($\theta = 0.1$, $n = 50$). The total height of each column indicates the probability that $\tilde{k}_A = \tilde{k}_B = 2$ and $\tilde{r}_{11}$ takes a value in the interval covered by the base of the column. The heights of the solid black columns indicate the estimated probability that $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 \leq 0.9$, $\tilde{q}_1 \leq 0.9$ and $\tilde{r}_{11}$ belongs to the indicated interval. For $\theta = 0.1$ and $n = 50$, the probabilities of $\tilde{k}_A = \tilde{k}_B = 2$ were estimated to be 0.085, 0.83, 0.084 and 0.083 for $C = 0$, $C = 5$, $C = 20$ and $C = \infty$, respectively. The probabilities of ($\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 \leq 0.9$ and $\tilde{q}_1 \leq 0.9$) are approximately 0.026, 0.023, 0.023 and 0.022 for $C = 0$, 5, 20 and ∞, respectively.

The distribution of $\tilde{D}'_{11}$ is shown in Figure 4. For $C = 0$ and $\tilde{k}_A = \tilde{k}_B = 2$, the absolute value of $\tilde{D}'_{11}$ is always 1. As C increases the probability of intermediate values of $\tilde{D}'_{11}$ increases, as shown in Figure 4. Even with C very large, the probability that $|\tilde{D}'_{11}|$ equals 1 is substantial. For $C = \infty$, $n = 50$ and conditional on $\tilde{k}_A = \tilde{k}_B = 2$, the probability that $|\tilde{D}'_{11}|$ equals 1 is 0.62. However, if one considers only samples with $\tilde{p}_1 \leq 0.9$ and $\tilde{q}_1 \leq 0.9$, then the conditional probability is 0.09.

For samples of size 50 and $\theta = 0.1$, estimated distributions of ($\tilde{D}_{11}$, $\tilde{p}_1$, $\tilde{q}_1$; $\tilde{k}_A = \tilde{k}_B = 2$) are shown in Figure 5 for six different ($\tilde{p}_1$, $\tilde{q}_1$) pairs and seven different recombination rates. Note that the vertical scale is not the same throughout the figure. Also, the horizontal scale is different for each specified ($\tilde{p}_1$, $\tilde{q}_1$). In each case the interval shown on the horizontal axis is from $D_a = -(1 - \tilde{p}_1)(1 - \tilde{q}_1)$ to $D_b = (1 - \tilde{p}_1)\tilde{q}_1$. Given $\tilde{p}_1$ and $\tilde{q}_1$, $\tilde{D}_{11}$ is necessarily in the closed interval $[D_a, D_b]$.

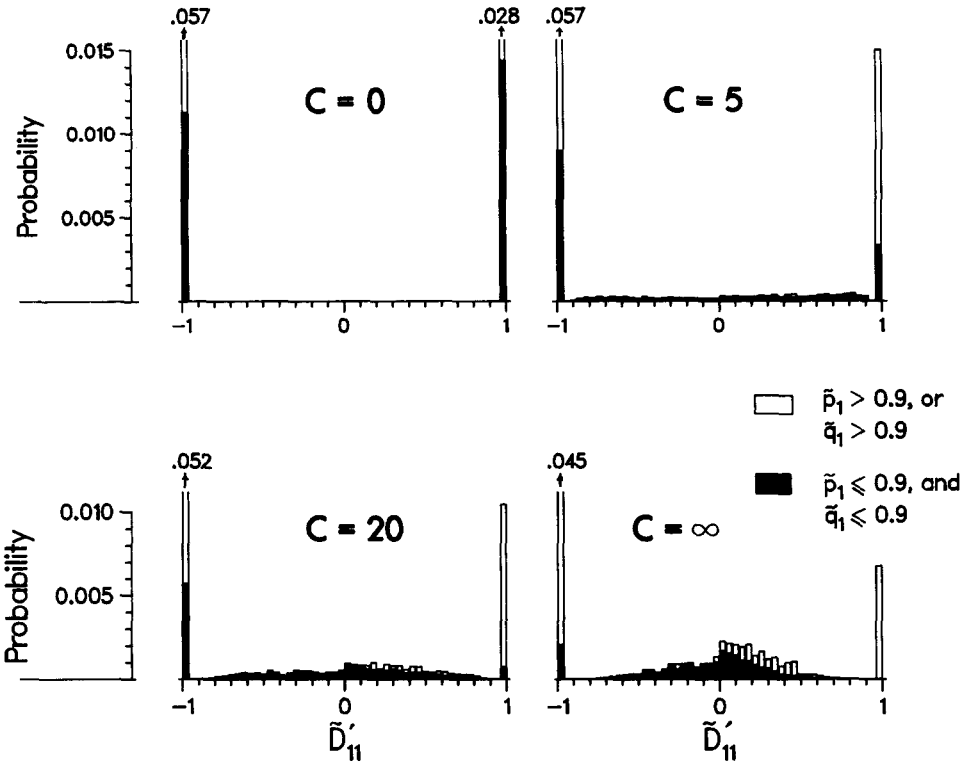

FIGURE 4.—Estimates of the distribution of $\tilde{D}'_{11}$ at four levels of recombination ($\theta = 0.1$, $n = 50$). The total height of each column indicates the probability that $\tilde{k}_A = \tilde{k}_B = 2$ and $\tilde{D}'_{11}$ takes a value in the interval covered by the base of the column. The heights of the solid black columns indicate the estimated probability that $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 \leq 0.9$, $\tilde{q}_1 \leq 0.9$ and $\tilde{D}'_{11}$ belongs to the indicated interval. For $\theta = 0.1$ and $n = 50$, the probabilities of $\tilde{k}_A = \tilde{k}_B = 2$ were estimated to be 0.085, 0.83, 0.084 and 0.083 for $C = 0$, $C = 5$, $C = 20$ and $C = \infty$, respectively. The probabilities of ($\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 \leq 0.9$ and $\tilde{q}_1 \leq 0.9$) are approximately 0.026, 0.023, 0.023 and 0.022 for $C = 0$, 5, 20 and ∞, respectively.

For $\tilde{p}_1 \leq 0.8$ and $\tilde{q}_1 \leq 0.8$, the distributions of $(\tilde{D}_{11} \mid \tilde{p}_1, \tilde{q}_1, \tilde{k}_A = \tilde{k}_B = 2)$ (which are easily obtained from the distributions in Figure 5) can be described as follows for different values of C. For $C = 0$, $\tilde{D}_{11}$ equals either D_a or D_b ($|\tilde{D}'_{11}| = 1$). For $C < 2$, the conditional distribution of $\tilde{D}_{11}$ remains strongly U shaped, with most of the probability mass at D_a and D_b. For $2 < C < 10$, the distribution of $\tilde{D}_{11}$ is fairly uniform over the interval $[D_a, D_b]$. For $C = 20$, the distribution is unimodal but has considerably higher variance than the distribution with $C = \infty$. For $C = 50$, the distribution of $\tilde{D}_{11}$ is close to the $C = \infty$ distribution but still has significantly more probability mass in the tails of the distributions than for $C = \infty$.

For some recombination rates, for example $C = 5$, the distribution of $(\tilde{D}_{11} \mid \tilde{p}_1, \tilde{q}_1, \tilde{k}_A = \tilde{k}_B = 2)$ appears to be fairly insensitive to the specified value of $(\tilde{p}_1, \tilde{q}_1)$. It is important to recall, however, that the horizontal scale is different for each $(\tilde{p}_1, \tilde{q}_1)$. Also, for tight linkage, there appear to be three

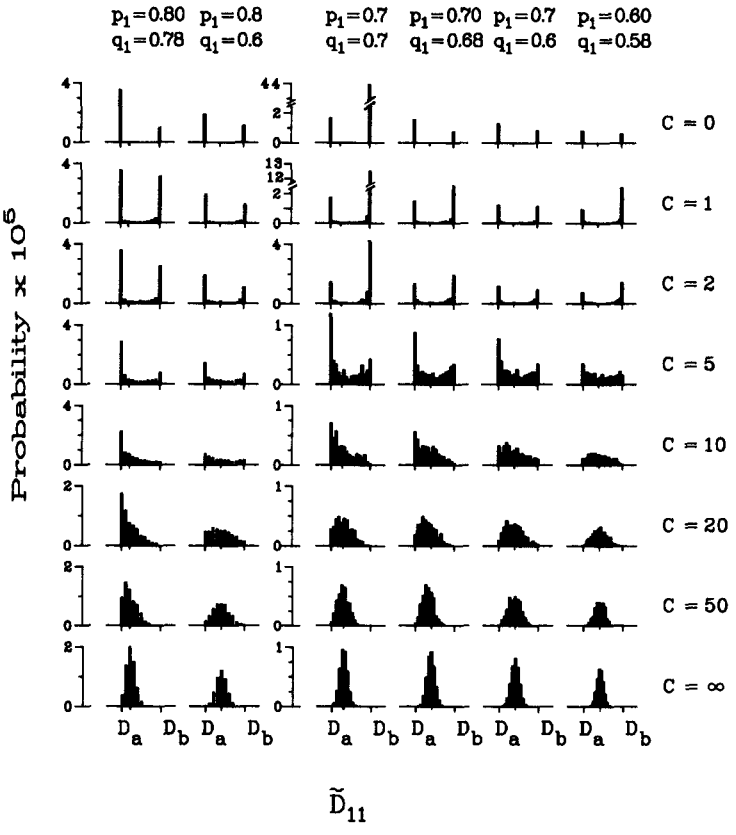

FIGURE 5.—Estimates of the joint distribution of $\tilde{D}_{11}$, $\tilde{p}_1$ and $\tilde{q}_1$ ($\theta = 0.1$, $n = 50$). Height of each column gives an estimate of the probability of ($\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 = p_1$, $\tilde{q}_2 = q_2$ and $D \in I$), where p_1 and q_2 are the frequencies indicated at the top of the figure, and where I is the interval at the base of the column. D_a is $-(1 - \tilde{p}_1)(1 - \tilde{q}_1)$, and D_b is $(1 - \tilde{p}_1)\tilde{q}_1$. Notice that both the horizontal and vertical scales differ from one histogram to the next. The histograms for $C = \infty$ are exact results using the sampling theory for one locus (EWENS 1972) to obtain the probability of the allele frequencies and assuming complete independence of loci for determining the distribution of the alleles on gametes.

situations that lead to quite distinct distributions of ($\tilde{D}_{11} \mid \tilde{p}_1, \tilde{q}_1, \tilde{k}_A = \tilde{k}_B = 2$), namely, $\tilde{p}_1$ equal to $\tilde{q}_1$, $\tilde{p}_1$ nearly equal to $\tilde{q}_1$ and $\tilde{p}_1$ quite different from $\tilde{q}_1$. These three cases are illustrated in Figure 5 by ($\tilde{p}_1$, $\tilde{q}_1$) equal to (0.7, 0.7), (0.7, 0.68) and (0.7, 0.6), respectively. When $\tilde{p}_1$ is not equal to $\tilde{q}_1$ the conditional expectation of $\tilde{D}_{11}$ is close to 0. For example, with ($\tilde{p}_1$, $\tilde{q}_1$) = (0.7, 0.6) and $C = 0$, $\tilde{D}_{11}$ equals -0.12 with probability 0.6, approximately, and $\tilde{D}_{11}$ equals 0.18 with probability 0.4 approximately, so the approximate conditional mean of $\tilde{D}$ is 0. When $\tilde{p}_1$ is nearly equal to $\tilde{q}_1$, the conditional expectation of $\tilde{D}_{11}$ is similarly close to 0 when $C = 0$. But when $\tilde{p}_1 = \tilde{q}_1$ and $C = 0$, the conditional expectation of $\tilde{D}_{11}$ is nearly equal to $\tilde{D}_b$, the maximum possible value of $\tilde{D}_{11}$ consistent with $\tilde{p}_1$ and $\tilde{q}_1$. For example, when ($\tilde{p}_1$, $\tilde{q}_1$) = (0.7, 0.7), the estimated conditional probabilities that $\tilde{D}_{11}$ equals -0.09 and 0.21 are 0.04

and 0.96, respectively, which implies an estimated conditional mean of $\tilde{D}_{11}$ equal to 0.20. For $\tilde{p}_1 = \tilde{q}_1$, the conditional expectation of $\tilde{D}_{11}$ decreases steadily as C increases. For $C = 5$, the expectation of $\tilde{D}_{11}$ conditional on $(\tilde{p}_1, \tilde{q}_1) = (0.7, 0.7)$ is approximately 0.03. For $\tilde{p}_1$ quite different from $\tilde{q}_1$, the conditional expectation of $\tilde{D}_{11}$ also decreases as C increases. But when $\tilde{p}_1$ is nearly equal to $\tilde{q}_1$, the probability that $\tilde{D}_{11} = D_b$ increases initially as C increases from 0. The result is that the conditional expectation of $\tilde{D}_{11}$ actually increases with increasing C. For $C = 0$ and $(\tilde{p}_1, \tilde{q}_1) = (0.7, 0.68)$, the conditional expectation of $\tilde{D}_{11}$ is 0.003, whereas for $C = 1$, the conditional expectation is approximately 0.09 (D_b is 0.204). This conditional expectation of $\tilde{D}_{11}$ does not increase indefinitely with increasing C, in fact, it begins decreasing when C equals approximately 1. For $C = 2$, this conditional expectation of $\tilde{D}_{11}$ is approximately 0.08; for $C = 5$, the expectation is about 0.03.

No cases in which $\tilde{p}_1 > 0.8$ and/or $\tilde{q}_1 > 0.8$ are shown in Figure 5, but results not shown indicate that the extreme values of $\tilde{D}_{11}$ (where $|\tilde{D}'_{11}| = 1$) can have substantial probabilities of occurring for these allele frequencies even if $C = \infty$.

In Figure 6, the conditional distribution of $\tilde{D}_{11}$ for $n = 20$, 50 and 100 are compared. For $C = 5$ and $C = 10$, the sample size has little effect. At higher levels of recombination, the larger samples differ in having lower probabilities for the extreme values of $\tilde{D}_{11}$.

Figure 7 shows that the distribution of $\tilde{D}_{11}$ conditional on the allele frequencies is quite independent of θ, at least for θ between 0.02 and 0.2.

DISCUSSION

Inferences about C: Several authors have estimated C or N using the statistic $\tilde{r}^2$ (LANGLEY 1977; LAURIE-AHLBERG and WEIR 1979; HILL 1981). To obtain an estimate of N, Langley assumed that $E(\tilde{r}^2) \approx 1/C + 1/n$, whereas Laurie-Ahlberg and Weir and Hill assumed that $E(\tilde{r}^2) \approx [(1 - c)^2 + c^2]/[2Nc(2 - c)] + 1/n$, which approximately equals $1/C + 1/n$ when $c \ll \frac{1}{2}$. (Throughout this paper, it has been assumed that c is small.) These assumed relationships between $E(\tilde{r}^2)$ and C clearly cannot apply when C is small, since $\tilde{r}$ is always less than or equal to 1. In addition, these relationships were derived for a two-locus model without mutation. The extent to which the behavior of models without mutation can be used to predict the behavior of models with mutation is unclear, although σ_d^2, a suggested approximation for $E(\tilde{r}^2)$ under the neutral model with mutation, is also approximately equal to $1/C$ for C large (KIMURA and OHTA 1971). Table 2 shows that neither σ_d^2 nor $1/C + 1/n$ is a good approximation for $E(\tilde{r}^2)$. However, it is common practice in survey studies of linkage disequilibrium to eliminate from consideration those loci that are nearly monomorphic (see, for example, LANGLEY, TOBARI and KOJIMA 1974). This means that conditional expectations, such as $E(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$, are needed for interpreting the observations. Table 2 shows that σ_d^2 is a fairly good approximation for $E(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$.

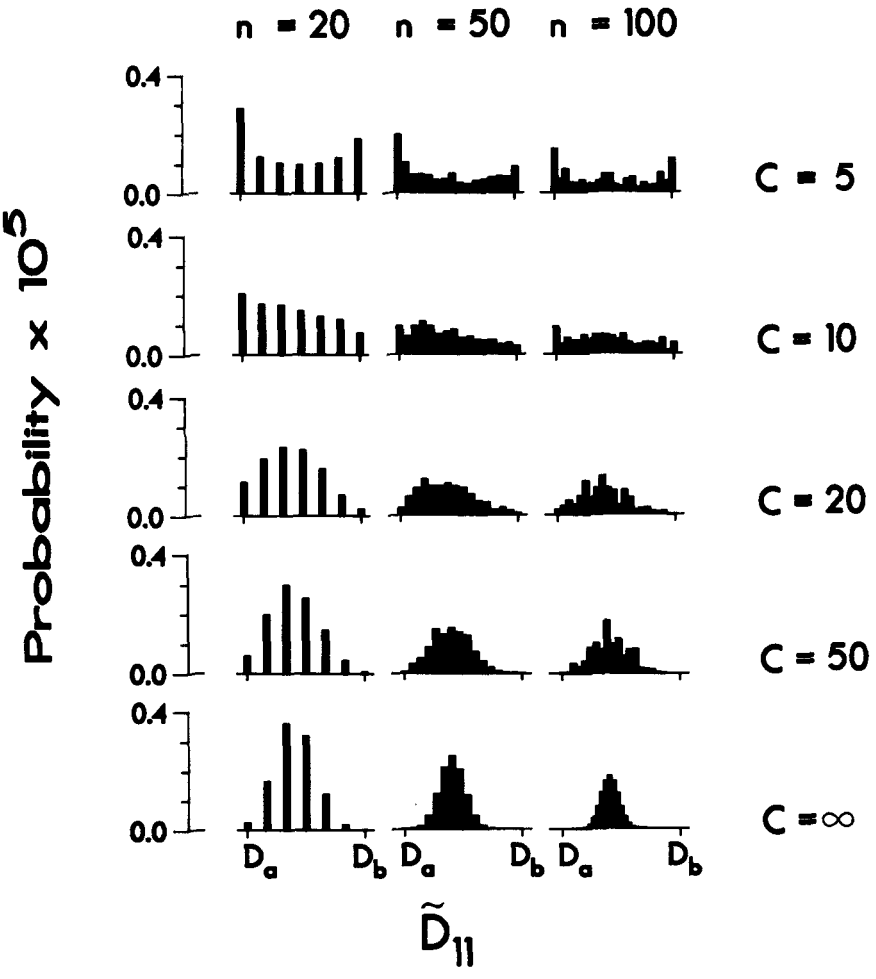

FIGURE 6.—Comparisons of the distribution of $\tilde{D}_{11}$ conditional on $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 = 0.7$ and $\tilde{q}_1 = 0.6$ for different sample sizes ($\theta = 0.1$). Actually, the distributions for $n = 100$, are averages of estimated distributions for $(\tilde{p}_1, \tilde{q}_1)$ equal to (0.7, 0.6), (0.7, 0.58), (0.7, 0.62), (0.68, 0.6) and (0.72, 0.6). This was done to obtain a more accurate estimate with a minimum of computer time.

Empirically, the best approximation to this conditional expectation appears to be:

$$E(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95) \approx \sigma_d^2 + 1/n, \qquad (6)$$

where σ_d^2 is calculated using equation (10) by HILL (1975), and with $\theta = 4N\mu$ set equal to 0.1, regardless of the true value of θ. Note that one does not need to know or estimate θ to estimate C using (6). Table 2 shows that this approximation works well for all cases examined, that is, for sample sizes of 50 to 200, for C between 0 and 20, and with θ between 0.02 and 0.4. The conditional expectation is equally well approximated by $1/C + 1/n$ for $C > 20$. That σ_d^2 is a good approximation for $E(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$ may be somewhat fortuitous, however, since other simulation results, shown in Fig-

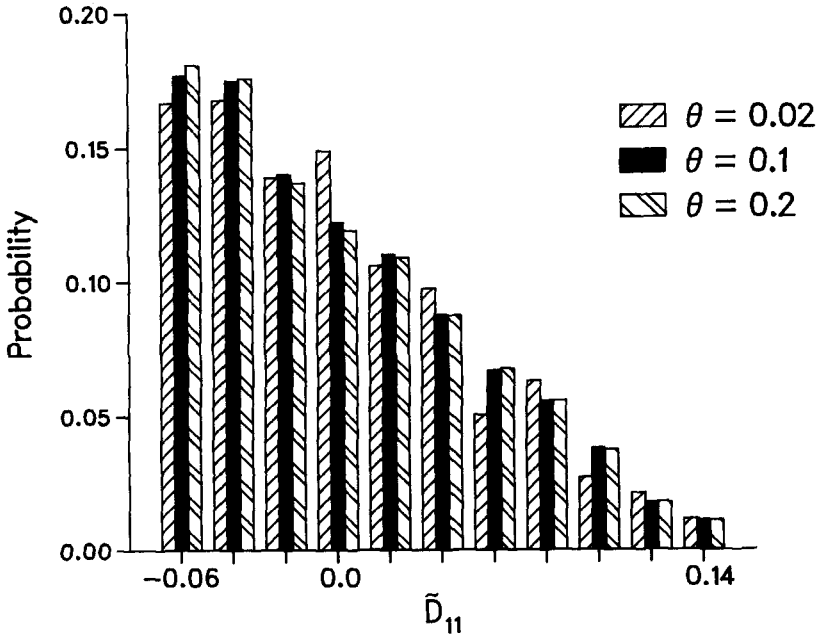

FIGURE 7.—Comparisons of the distribution of $\tilde{D}_{11}$ conditional on $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 = 0.8$ and $\tilde{q}_1 = 0.7$ for different values of $\theta(n = 50)$. Evidently there is little dependence of this conditional distribution on θ. Simulations with other values of $\tilde{p}_1$ and $\tilde{q}_1$ indicate that this result holds more generally.

ure 8, indicate σ_d^2 is not a good approximation for $E(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.9, \tilde{q}_1 \leq 0.9)$ or $E(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.975, \tilde{q}_1 \leq 0.975)$ when linkage is tight.

Given an adequate approximation to the conditional mean of $\tilde{r}^2$, equation (6), we now address the question: Assuming that the neutral model is correct, how much information about C can be obtained from a single sample for which $\tilde{D}_{11}$, $\tilde{p}_1$ and $\tilde{q}_1$ are measured? Examination of Figure 3 shows that $\tilde{r}_{11}$ alone, without considering it jointly with $\tilde{p}_1$ and $\tilde{q}_1$, will not generally be very informative about C. This is clear since there is a large overlap of the distribution of $\tilde{r}_{11}$ with $C = 0$ and the distribution with $C = \infty$. Small negative values of $\tilde{r}_{11}$ are very likely for all values of C. Similarly, Figures 2 and 4 show that $\tilde{D}_{11}$ and $\tilde{D}'_{11}$ are not very informative about C when the frequencies, $\tilde{p}_1$ and $\tilde{q}_1$, are not considered. Now consider the use of the joint distribution of $(\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1; \tilde{k}_A = \tilde{k}_B = 2)$ for making inferences about C. Examination of Figure 5 reveals that any observed value of $\tilde{D}_{11}$ is compatible with values of C between 2 and 10. For C in this interval, $\tilde{D}_{11}$ is very broadly distributed over the interval $[D_a, D_b]$. (Recall, $D_a = -(1 - \tilde{p}_1)(1 - \tilde{q}_1)$ is the minimum possible value of $\tilde{D}_{11}$ given $\tilde{p}_1$ and $\tilde{q}_1$. And $D_b = (1 - \tilde{p}_1)\tilde{q}_1$ is the maximum possible value of $\tilde{D}_{11}$ given $\tilde{p}_1$ and $\tilde{q}_1$.) However, very small and very large values of C can be ruled out by certain observed values of $\tilde{D}_{11}$. If the observed value of $\tilde{D}_{11}$ were near D_a or D_b, then one could rule out very large values of C. For example, if one observed $(\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1) = (0.164, 0.7, 0.68)$ in a sample of 50 gametes, one could conclude that with high probability $C < 20$. This is because, when $C =$

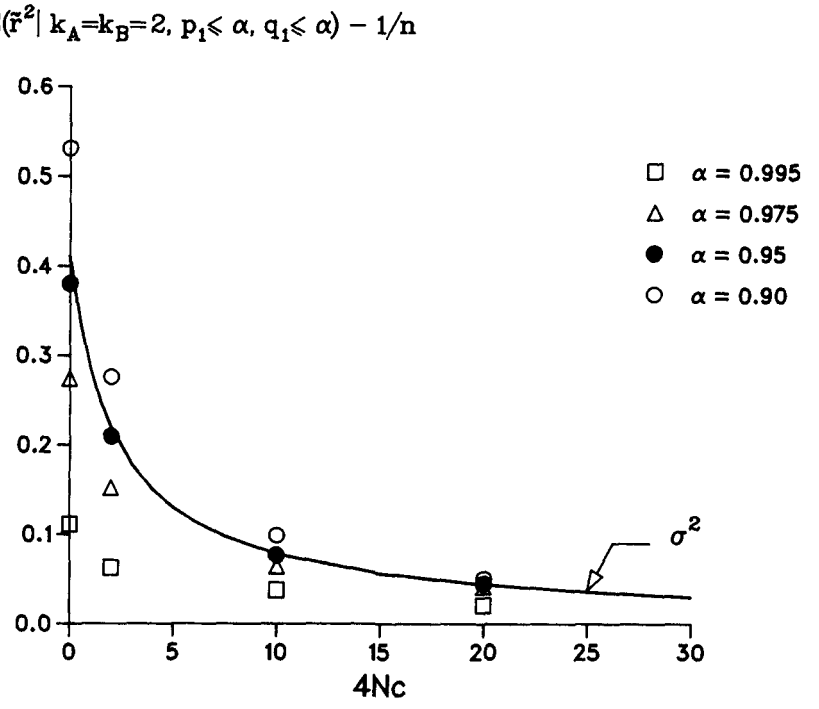

FIGURE 8.—A comparison of $E(\tilde{r}^2 \mid \tilde{k}_A = \tilde{k}_B = 2, \tilde{p}_1 \leq \alpha, \tilde{q}_1 \leq \alpha) - 1/n$ and σ_d^2 for various values of α ($\theta = 0.1$, $n = 200$). The curve is σ_d^2 calculated using equation (10) from HILL (1975).

20, the probability that $|\tilde{D}_{11}|$ is greater than or equal to 0.164, given $\tilde{p}_1 = 0.7$ and $\tilde{q} = 0.68$, is approximately 0.01, and, for all C greater than 20, the probability is even smaller. Recall that the distribution of $\tilde{D}_{11}$ given $\tilde{p}_1$ and $\tilde{q}_1$ is essentially independent of θ, so these conclusions can be drawn without precise knowledge of the value of θ. If a value of $\tilde{D}_{11}$ near 0 were observed, one could rule out very small values of C. For example, the probability of $|\tilde{D}_{11}| < 0.04$ is approximately 0.03 for $C = 1$, given that $\tilde{p}_1 = 0.7$ and $\tilde{q}_1 = 0.68$. For all C less than 1, it appears that the probability is smaller. Evidently, if one obtained a sample such that $(\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1) = (-0.036, 0.7, 0.68)$, a conclusion that C is greater than 1 is justified. In summary, a single observation, $(\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1)$ will typically be sufficient to establish a rather large upper bound on C (*e.g.*, 20) and/or a rather small lower bound on C (*e.g.*, 1). No more precise conclusions concerning C can be drawn.

The results in Figure 5 also permit one to use an observation of $\tilde{D}_{11}$, $\tilde{p}_1$ and $\tilde{q}_1$ to obtain an approximate maximum likelihood estimate of C or, if c is known, of N, the population size. To illustrate this, the estimated probability of $(\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1) = (0.08, 0.8, 0.6)$ as a function of C is shown in Figure 9. Seven of the points on Figure 9 are taken directly from Figure 5. The rest of the points are from simulations results not shown in Figure 5. The maximum likelihood estimate of C based on the observation $(\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1) = (0.08, 0.8, 0.6)$ is approximately 12. For many observations the maximum likelihood es-

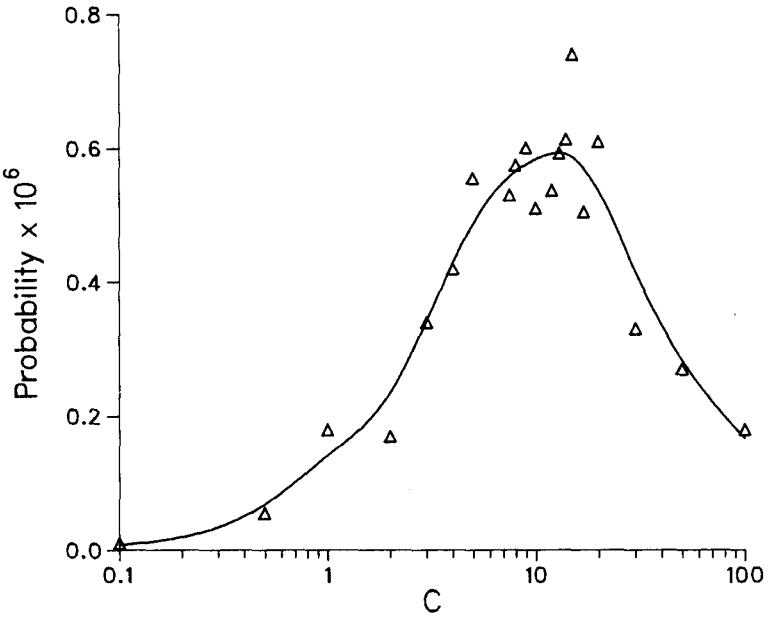

FIGURE 9.—Estimates of the probability of ($D_{11} = 0.08$; $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 = 0.8$, $\tilde{q}_1 = 0.6$) as a function of C are shown by triangles (△) ($\theta = 0.1$, $n = 50$). The curve is an arbitrary smooth curve. An approximate maximum likelihood estimate of C based on a sample for which ($\tilde{D}_{11} = 0.08$; $\tilde{k}_A = \tilde{k}_B = 2$, $\tilde{p}_1 = 0.8$, $\tilde{q}_1 = 0.6$) is 12. Evidently, $C = 3$ and $C = 50$ are not a great deal less likely.

timate of C would be an extreme value, either 0 or infinity. Interestingly, the maximum likelihood estimate of C is often, but not always, 0 when $\tilde{D}_{11} = D_b$, the highest possible linkage disequilibrium consistent with the allele frequencies. As shown in Figure 5, for ($\tilde{p}_1$, $\tilde{q}_1$) equal to (0.8, 0.78), (0.7, 0.68) or (0.6, 0.58) and $\tilde{D}_{11} = D_b$, the maximum possible value of $\tilde{D}_{11}$, given the allele frequencies, the probability of the sample is greater with $C = 1$ than with $C = 0$.

A maximum likelihood estimate based on one pair of loci in a single sample is clearly not very informative. In the example shown in Figure 9, the maximum likelihood estimate is approximately 12, but the likelihood of $C = 3$ and $C = 50$ are not a great deal smaller. However, if several pairs of loci were examined and each pair of loci were distant enough from the others that independence could be assumed, then a maximum likelihood estimate based on all of the pairs of loci could be obtained. Such an estimate would be very useful, especially if certain asymptotic properties of maximum likelihood estimators could be invoked.

Hypothesis testing: Now consider the equilibrium neutral model as a null hypothesis that we wish to test against an alternative selective hypothesis. Suppose that C is known and that the alternative hypothesis is a selective model under which strong linkage disequilibrium is expected. First, consider a test based on the distribution of $|\tilde{D}_{11}|$ conditional on both $\tilde{k}_A = \tilde{k}_B = 2$ and the

observed allele frequencies. If the distributions of $\tilde{p}_1$ and $\tilde{q}_1$ were very different under the null and alternative hypotheses, this would not be a particularly powerful test, but if one wishes to base the test only on the amount of association between alleles without regard to whether the allele frequencies themselves are more likely under one hypothesis than the other, then this test would be appropriate. For what values of C can we expect to be able to reject the null hypothesis on the basis that $|\tilde{D}_{11}|$ is too large to be compatible with the null hypothesis? Without specifying more precisely the alternative hypothesis, the power of the test cannot be determined but some general observations can be made. Figure 5 shows that for $C < 10$, rejection of the neutral model on the grounds that $|\tilde{D}_{11}|$ is too large is very unlikely regardless of the alternative hypothesis. For example, with $C = 5$ and conditional on $p_1 = 0.7$, and $q_1 = 0.6$, the probability that $|\tilde{D}_{11}| > 0.184$ is approximately 0.16. The maximum value of $\tilde{D}_{11}$ given these allele frequencies is 0.204. For $C = 10$, the neutral model is likely to be rejected under the alternative model only if the alternative model is such that $\tilde{D}_{11}$ is expected to take the highest possible value, given the allele frequencies, and the allele frequencies are not too far from 0.5. Even at $C = 20$, if under the alternative hypothesis $\tilde{D}_{11}$ is expected to be large in absolute value but negative in sign, the neutral model will not be rejectable unless the allele frequencies are close to 0.5. For $C = 50$, Figure 5 shows that the distribution of $\tilde{D}_{11}$ under the neutral model is getting close to the no-linkage distribution. In this case, if strong linkage disequilibrium is expected under the alternative hypothesis, the power of a test may be substantial. However, the critical values for the test when $C = 50$ are still substantially different from the critical values under the $C = \infty$ neutral model. For example, with $\tilde{p}_1 = 0.7$ and $\tilde{q}_1 = 0.68$, the probability that $|\tilde{D}_{11}|$ is greater than 0.07 is 0.019 for $C = \infty$, whereas for $C = 50$, this probability is 0.096. Under the neutral hypothesis with $C = \infty$, the power of the tests for different alternative hypotheses has been discussed by BROWN (1975).

Conclusions: The unconditional expectation of $\tilde{r}^2$ is not well approximated by σ_d^2 when linkage is tight. However, equation (6) gives a useful approximation for $E(\tilde{r}^2 \mid k_A = \tilde{k}_B = 2, \tilde{p}_1 \leq 0.95, \tilde{q}_1 \leq 0.95)$ that applies for a large range of sample sizes, mutation rates and values of C. For $C > 20$, the approximation $E(\tilde{r}^2 \mid k_A = k_B = 2; \tilde{p}_1 \leq 0.95), \tilde{q} \leq 0.95) \approx 1/C + 1/n$ is quite good. Although this conditional mean of $\tilde{r}^2$ can be approximated, the distribution of $\tilde{r}^2$ about that mean is shown to have poor statistical properties. In particular $(\tilde{r}_{11} \mid \tilde{k}_A = \tilde{k}_B = 2)$ is shown to have large variance and to be bimodal for tight linkage. This motivates the study of the joint distribution of $\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1$.

Using the estimated joint distribution of $\tilde{D}_{11}, \tilde{p}_1, \tilde{q}_1$, one can obtain an approximate maximum likelihood estimate of C. The distribution of $\tilde{D}_{11}$, conditional on $\tilde{p}_1$ and $\tilde{q}_1$, is shown to be very insensitive to θ. It is shown that $E(\tilde{D}_{11} \mid \tilde{p}_1, \tilde{q}_1)$ is quite large when C is small and $\tilde{p}_1 = \tilde{q}_1$. Also, if $\tilde{p}_1$ is nearly but not exactly equal to $\tilde{q}_1$, then $E(\tilde{D}_{11} \mid \tilde{p}_1, \tilde{q}_1)$ is larger for $C = 1$ than for $C = 0$.

This work benefited from discussions with C. AQUADRO and C. STROBECK. N. KAPLAN and C. AQUADRO provided valuable criticism of an earlier version of the manuscript. G. B. GOLDING

generously provided unpublished results which were invaluable for checking the computer programs.

LITERATURE CITED

BROWN, A. H. D., 1975 Sample sizes required to detect linkage disequilibrium between two and three loci. Theor. Pop. Biol. **8:** 184–201.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Pop. Biol. **3:** 87–112.

EWENS, W. J., 1979 *Mathematical Population Genetics.* Springer-Verlag, New York.

FRANKLIN, I. and R. C. LEWONTIN, 1970 Is the gene the unit of selection? Genetics **65:** 701–734.

GOLDING, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108:** 257–274.

GOLDING, G. B. and C. STROBECK, 1983 Two-locus, fourth order gene frequency moments: implications for the variance of squared linkage disequilibrium and the variance of homozygosity. Theor. Pop. Biol. **24:** 173–191.

GRIFFITHS, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Pop. Biol. **19:** 169–186.

HILL, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Pop. Biol. **8:** 117–126.

HILL, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. **38:** 209–216.

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

KARLIN, S. and J. L. MCGREGOR, 1968 Rates and probabilities of fixation for two locus random mating finite populations without selection. Genetics **58:** 141–159.

KIMURA, M. and T. OHTA, 1971 *Theoretical Aspects of Population Genetics.* Princeton University Press, Princeton, New Jersey.

LANGLEY, C. H., 1977 Nonrandom associations between allozymes in natural populations of Drosophila melanogaster. pp. 265–273. In: *Lecture Notes in Biomathematics,* Vol. 19: Measuring Selection in Natural Populations, Edited by F. B. CHRISTIANSEN and T. M. FENCHEL. Springer-Verlag, New York.

LANGLEY, C. H., Y. N. TOBARI and K. KOJIMA, 1974 Linkage disequilibrium in natural populations of *Drosophila melanogaster.* Genetics **78:** 921–936.

LAURIE-AHLBERG, C. C. and B. S. WEIR, 1979 Allozymic variation and linkage disequilibrium in some laboratory populations of *Drosophila melanogaster.* Genetics **92:** 1295–1314.

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

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York.

MARUYAMA, T., 1982 Stochastic integrals and their application to population genetics. pp. 151–166. In: *Molecular Evolution, Protein Polymorphism and the Neutral Theory,* Edited by M. Kimura. Springer-Verlag, Berlin.

OHTA, T. and M. KIMURA, 1969 Linkage disequilibrium due to random genetic drift. Genet. Res. **13:** 47–55.

OHTA, T. and M. KIMURA, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68:** 571–580.

STROBECK, C. and K. MORGAN, 1978 The effect of intragenic recombination on the number of alleles in a finite population. Genetics **88:** 829–844.

TAKAHATA, N., 1982 Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. Genet. Res. **39:** 63–77.

Communicating editor: M. NEI

APPENDIX

I derive here an expression for the expectation of $\hat{D}^2$ in a sample under the two-locus neutral infinite allele model.

Define the following five identity coefficients:

$$\Phi_A = E\left(\sum_i p_i^2\right) \tag{A1}$$

$$\Phi_B = E\left(\sum_i q_i^2\right) \tag{A2}$$

$$\Phi_{AB} = E\left(\sum_i \sum_j f_{ij}^2\right) \tag{A3}$$

$$\Gamma_{AB} = E\left(\sum_i \sum_j p_i q_j f_{ij}\right) \tag{A4}$$

$$\Delta_{AB} = E\left(\sum_i \sum_j p_i^2 q_j^2\right) \tag{A5}$$

ϕ_A is the probability that two gametes drawn randomly with replacement from the population are identical at locus A. ϕ_B is the probability that two gametes drawn randomly with replacement from the population are identical at locus B. In what follows, it will be assumed that $\phi_A = \phi_B$. ϕ_{AB} is the probability that two gametes drawn randomly with replacement from the population are identical at both locus A and locus B. Γ_{AB} is the probability that, when three gametes are drawn, the second is identical with the first at the A locus, and the third is identical with the first at the B locus. And finally, Δ_{AB} is the probability that, when four gametes are drawn, the first is identical with the second at the A locus and the third is identical with the fourth at the B locus.

Notice that D^2 can be written as:

$$\begin{aligned} E(D^2) &= E\left(\sum_i \sum_j [f_{ij} - p_i q_j]^2\right) \\ &= \phi_{AB} - 2\Gamma_{AB} + \Delta_{AB} \end{aligned} \tag{A6}$$

(STROBECK and MORGAN 1978). STROBECK and MORGAN (1978) provide formulas for each of these coefficients. Now, suppose a sample of gametes is drawn from the population. If the population size is large enough, it does not matter whether the sampling is done with replacement or, as is typical, without replacement. In what follows, it will be assumed that the probability that two distinct gametes in the sample are identical is ϕ_{AB}. This requires that the sampling be done with replacement or that the population size be large enough that f_{ij} is essentially equal to $f_{ij} - 1/N$. As with the population statistic, we can write $\hat{D}^2$ in terms of identity coefficients:

$$E(\hat{D}^2) = \tilde{\phi}_{AB} - 2\tilde{\Gamma}_{AB} + \tilde{\Delta}_{AB}, \tag{A7}$$

where the coefficients with a tilde ($\sim$) are defined in the obvious way, in terms of the sample frequencies of alleles and gametes. These identity coefficients can be interpreted in the same way as the population coefficient except that sampling is from the sample instead of from the population. For example, $\tilde{\phi}_{AB}$ is the probability that two gametes drawn at random with replacement from the sample are identical at both locus A and locus B. We now derive expressions for these

sample identity coefficients in terms of the population coefficients. What is the probability that two gametes drawn with replacement from the sample are identical at both loci? With probability $1/n$ the same gamete will be drawn from the sample twice, and with probability $1 - 1/n$ two distinct gametes in the sample will be drawn. This implies that

$$\tilde{\phi}_{AB} = 1/n + (1 - 1/n)\phi_{AB}. \tag{A8}$$

Similarly, with probability $(1 - 1/n)(1 - 2/n)$ three gametes drawn from the sample with replacement are all distinct gametes in the sample, and then the probability that the second is identical with the first at the locus A and the third is identical with the first at locus B is just Γ_{AB}. With probability $(1/n)(1 - 1/n)$ the second gamete drawn is just a resampling of the first gamete and the third gamete is a distinct gamete from the sample. If the gametes were drawn in this way, the probability of the second being identical with the first at locus A is 1, and the probability of the third gamete being identical with the first at the B locus is just ϕ_B. Considering the probability of the first and third sampled gametes being just resampling of the same gamete, the probability that second and third gametes are resamplings of the same gamete, and also the probability of the same gamete being sampled three times from the sample, we can write

$$\tilde{\Gamma}_{AB} = \frac{(n-1)(n-2)}{n^2}\Gamma_{aB} + \frac{2}{n}\left(1 - \frac{1}{n}\right)\phi_A + \frac{1}{n}\left(1 - \frac{1}{n}\right)\phi_{AB} + \frac{1}{n^2}. \tag{A9}$$

Δ_{AB} is more complicated since it involves sampling four gametes from the sample, but the same reasoning leads to

$$\tilde{\Delta}_{AB} = \frac{(n-1)(n-2)(n-3)}{n^3}\Delta_{AB} + \frac{2(n-1)(n-2)}{n^3}[\phi_A + 2\Gamma_{AB}] + \frac{2(n-1)[2\phi_A + \phi_{AB}]}{n^3} + \frac{1}{n^2}. \tag{A10}$$

By substituting into (A8), (A9) and (A10) the expressions for ϕ_{AB}, Γ_{AB} and Δ_{AB} provided by STROBECK and MORGAN (1978) we obtain expressions for the expectations of the sample coefficients, which can then be substituted into (A7) to obtain $E(\tilde{D}^2)$.