The Detection of Linkage Disequilibrium in Molecular Sequence Data

R. C. Lewontin

Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138

Manuscript received July 15, 1994
Accepted for publication January 21, 1995

ABSTRACT

Studies of genetic variation in natural populations at the sequence level usually show that most polymorphic sites are very asymmetrical in allele frequencies, with the rarer allele at a site near fixation. When the rarer allele at a site is present only a few times in the sample, say below five representatives, it becomes very difficult to detect linkage disequilibrium between sites from tests of association. This is a consequence of the numerical properties of even the most powerful test of association, Fisher's exact test. Sites with fewer than five representatives in the sample should be excluded from association tests, but this generally leaves few site pairs eligible for testing. A test for overall linkage disequilibrium, based on the sign of the observed linkage disequilibria, is derived which can use all the data. It is shown that more power can be achieved by increasing the length of sequence determined than by increasing the number of genomes sampled for the same total work.

IN the study of molecular polymorphism, it has become a common practice to look for evidence of linkage associations between polymorphic sites along the sequence. The presence and pattern of nonrandom associations is, in turn, taken as evidence of selective and historical events forming the variation. Were polymorphism in a population very ancient and not under the influence of selection, we would expect polymorphic sites to be in linkage equilibrium with each other, while the recent spread of haplotypes through selection or migration, or the longtime maintenance of haplotypes by some form of balancing selection, will result in linkage disequilibria among sites. The usual practice is to calculate all pairwise linkage disequilibria among all polymorphic sites, test these for statistical significance, and make inferences from the pattern (or lack of pattern) along the sequence.

The polymorphism at a site consists almost always of only two alternatives, say *0* and *1*, with allele frequencies p_1,q_1 and p_2,q_2 at the two sites, respectively. Denoting the frequencies of the four haplotypic combinations *00*, *01*, *10*, and *11*, as g_{00}, g_{01}, g_{10}, and g_{11}, the usual *measure* of linkage disequilibrium is D' (LEWONTIN 1964):

$$D' = \frac{D}{D_{\max}} = \begin{cases} \dfrac{g_{11}g_{00} - g_{01}g_{10}}{\min\ (p_1q_2;\ p_2q_1)} & D \text{ pos (coupling)} \\ \dfrac{g_{11}g_{00} - g_{01}g_{10}}{\min\ (p_1q_1;\ p_2q_2)} & D \text{ neg (repulsion)} \end{cases}$$

The statistical test for the significance of D' is the usual test of association in 2×2 tables, either chi-square with one degree of freedom, or more usually, the most powerful Fisher's exact test. The question then arises whether the most powerful test indeed has enough power to detect linkage association in the samples usual in molecular population genetics. A number of papers have addressed the power of tests of association to detect linkages (BROWN 1975; FU and ARNOLD 1992; ZAPATA and ALVAREZ 1993), and all have concurred that quite large sample sizes, larger than are usual in molecular population genetic studies, are needed to detect moderate linkage disequilibria, especially when allele frequencies at the sites diverge markedly from 0.5. These studies are all cast in terms of the proportions p and q of the alleles at the sites, and the value of the linkage disequilibrium parameter D or its normalized form, D'. Such discussions in terms of proportions, while entirely correct, miss an essential feature of molecular population data, illustrated by Figure 1.

Figure 1, a and b, shows the distribution of allele frequencies in two recent studies of polymorphism (RILEY *et al.* 1989; SCHAEFFER and MILLER 1993), given in terms of k, the *absolute* number of an alternative allele, rather than the more usual relative frequency, $p = k/N$. The SCHAEFFER and MILLER study detected 359 polymorphic sites by DNA sequencing of a 3.5-kb region in a sample of 99 genomes. The RILEY *et al.* study was a four-cutter restriction enzyme study that showed 78 polymorphic restriction sites in a 4.6-kb region in a sample of 58 genomes. The abscissa in each figure gives the absolute number of times the rarer of the alleles at a site was seen in the sample, while the ordinate shows the proportion of all polymorphic sites that had this particular allele frequency. What is immediately striking about these data, and what is characteristic of virtually all sequence polymorphic data, is that a very large fraction of all polymorphic sites are represented in the data by singleton, or only a very few, copies of the rare allele.

Author e-mail: dick@mcz.harvard.edu

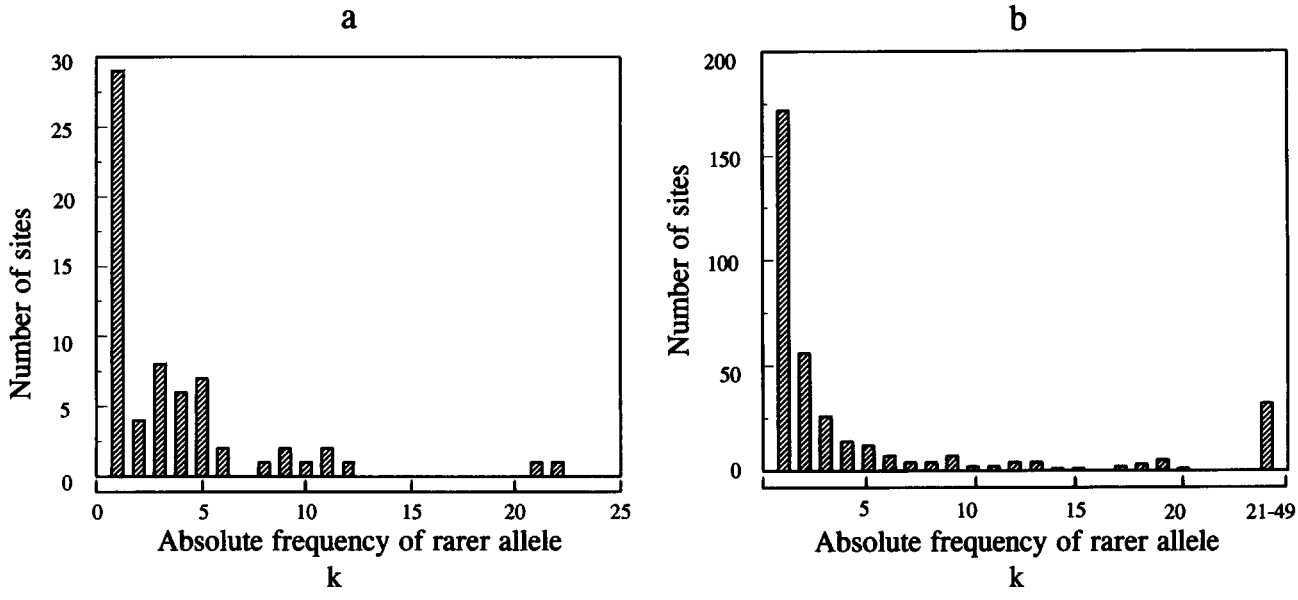

FIGURE 1.—Distribution of absolute number of copies of the rarer allele in two studies. Abscissa: number of copies, k, of the rarer allele at a site; ordinate: number of polymorphic sites with this copy number. (a) Data of RILEY *et al.* 1989. (b) Data of SCHAEFFER and MILLER.

In the very large sample of haplotypes studied by SCHAEFFER and MILLER, 75% of all polymorphic sites had the rarer allele represented four or fewer times. In the RILEY *et al.* data, the equivalent figure is 69%. This observation leads us to ask about linkage disequilibrium measures like D' and especially about tests of association in 2 × 2 tables when the *absolute* numbers in the margins of the table are very small. The results turn out to be rather unsettling.

The problem of tests of association in 2 × 2 tables when the marginal frequencies are as small as those observed in molecular polymorphism studies has no really satisfactory solution. When the true frequency of the rarer allele at one or both of the loci is <0.05, then the usual chi-square test, chi-square with Yates correction, and Fisher's exact test (FET) are all very conservative. That is, they give significant results less often than the nominal significance level probability (UPTON 1982; OVERALL *et al.* 1987; D'AGOSTINO *et al.* 1988), so there is no clear choice to be made among them on this basis. While it is sometimes said that FET applies only to contingency tables with nonsampled fixed marginals, this is incorrect. FET was derived on the assumption that the marginals are randomly sampled, and the test gives the exact probability of the table filling *conditional* on the observed marginals. Thus, FET avoids the necessity of estimating the true marginal *p*s as is done in chi-square and likelihood ratio tests. It is not usually noted that when the true marginal *p*s are small, the observed marginals are upwardly biased estimates of these *p*s because samples with 0 in a marginal do not get included in the data to be tested and these can be a substantial fraction of samples when marginal *p*s are small. Indeed, for an observed absolute marginal frequency of 1, there is no unbiased estimator of p when 0 marginals are left out. Finally, it was proved by TOCHER (1950) that FET is the uniformly most powerful test whether marginals are fixed or sampled. (See also KENDALL and STUART 1979.) Thus, in this paper, FET is used to demonstrate the problem.

GENERAL FORM OF THE OBSERVATIONS

We have a sample of N genomes in which some sites are polymorphic. For simplicity we assume, as is nearly always true, that there are two alternative allelic forms at each site, and we label the rarer allele *0* and the more common allele *1*. At any polymorphic site there are then 1, 2, 3, . . . , $N/2$ copies of the *0* allele. Let us consider pairwise combinations of polymorphic sites, and in each pair we will designate the site whose *0* allele is less frequent than the *0* allele at the other site as the *focal* site. The focal site will have k copies of allele *0* (and $N - k$ copies of allele *1*), while the other site will have m and n copies of allele *0* and *1*, respectively ($m + n = N$). By our convention of naming alleles,

$$k \leq m \leq n.$$

With these conventions of designating the rarer allele at each site and the focal site, we can now represent the data for any pair of sites as in Table 1. It then follows that

$$D' = \begin{cases} \dfrac{(m+n)\,j - mk}{nk} & \text{if } D \text{ is positive} \\ \dfrac{(m+n)\,j - mk}{mk} & \text{if } D \text{ is negative} \end{cases} \qquad (1)$$

TABLE 1

General form of the 2 × 2 association table

		Focal site		
		0	*1*	
Nonfocal site	*0*	j	$m - j$	m
	1	$k - j$	$n - k + j$	n
		k	$N - k$	N

With fixed marginal allele frequencies k, $N - k$, m, and n there are a number of possible fillings of the table. To concentrate our attention on the focal site, we will designate a filling of the table as $(j, k - j)$, although, of course, the other two cells of the table change as j changes. For example, if $k = 3$, the possible fillings of the table are designated as (3,0), (2,1), (1,2) and (0,3), corresponding to the four possible outcomes from complete coupling to complete repulsion of the rare alleles at the two loci. In general, there are exactly $k + 1$ possible fillings, irrespective of the values of m and n.

We designate the probability of a given random filling of the table under the null hypothesis of no linkage disequilibrium in the population as $P(j,k - j)$ and it is easily verified from the standard argument of FET that, given the marginals m, n, k, and $N - k$

$$P(j, k - j) = \frac{m!\,n!\,k!\,(N - k)!}{j!(k - j)!\ (m - j)!\ (n - k + j)!N!} \quad (2)$$

Note that only the denominator is a function of the actual filling.

Finally, FET probability is the combined probability of the observed filling of the table and any more extreme fillings. That is,

$$F(j, k - j) = \sum_{i=j}^{k} P(i, k - i) \quad \textit{for D pos} \quad (3a)$$

or

$$F(j, k - j) = \sum_{i=0}^{j} P(i, k - i) \quad \textit{for D neg} \quad (3b)$$

for one-sided tests of the null hypothesis of no linkage disequilibrium in the same direction as the observed deviation. For a test that allows disequilibria that may be either positive or negative, the probability in (2) is summed for all fillings that give as extreme or more extreme deviations from expectation than the observed filling, in either direction.

APPLICATION TO RARE ALLELES

Let us now consider what happens when the general formulas are applied to cases when k is very small.

$k = 1$: Most (but not all) studies have excluded the cases of singleton polymorphisms, because it is obvious that very little information can be contained in such cases. We begin with this case because it illustrates the general problems and the method for investigating them.

When the focal site has a singleton variant and $N - 1$ copies of the consensus allele, there are only two possible fillings of the table, (1,0) and (0,1). From Equations 1–3 above, we get immediately that

$$\text{Filling (1,0):} \quad D' = +1, \quad P(1,0) = F(1,0) = m/N$$

$$\text{Filling (0,1):} \quad D' = -1, \quad P(0,1) = F(0,1) = n/N$$

That is, D' *must* be either +1 or −1. Moreover the repulsion equilibrium can *never* be significant because $n/N > 0.5$ by our convention of naming alleles. The coupling configuration *could* be significant in a one-sided test if N is large enough, but only if the nonfocal site is also a site with a very asymmetrical allele frequency. So, if $N < 20$, the test will never be significant at the 5% level, while for $N = 100$, say, the nonfocal site must have $m \leq 5$ for any significance at a conventional level. But the overwhelming likelihood for any population genetic process, except for some heterotic cases with epistasis or a large increase in frequency of a previously rare haplotype, is that the rare allele at any site will be associated with the common allele at the other site, *i.e.*, that D' is negative. So a one-sided test will almost never be significant and a two-sided test, obviously, has a probability identically equal to 1.0 when $k = 1$.

The point of the foregoing demonstration is not that FET is *conservative* for $k = 1$, or that it has low power, both of which are issues of the probability distribution of test results, but that it is *numerically impossible* to achieve a significant test, *no matter what the sample size* when D' is negative, and only if N is greater than the reciprocal of the significance level when D' is positive. It should not be supposed that this numerical problem is unique to the FET. On the contrary, the usual chi-square test, with r without Yates correction for continuity is even more badly behaved for $k = 1$ and especially for small m. For

D negative (uncorrected):

$$\chi^2 = \frac{Nm}{(N - 1)\ (N - m)} = \frac{m}{(N - m)}$$

D negative (corrected): inapplicable

D positive (uncorrected):

$$\chi^2 = \frac{N(N - m)}{(N - 1)m} \simeq \frac{(N - m)}{m}$$

D positive (corrected):

$$\chi^2 = \frac{N\left(\frac{N}{2} - m\right)^2}{(N - 1)\ (N - m)m} \simeq \frac{(N - 2m)^2}{4(N - m)m}$$

So, negative D's can never be significant because chi-square with one degree for freedom will be less than 1, and for most value of m in the sample will, in fact, be quite small. Any overall test for linkage disequilibrium that sums over the chi-squares for a set of pairwise tests in which there are many pairs involving singletons will thus be strongly biased toward nonsignificance irrespective of how much true negative disequilibrium exists. In contrast, positive disequilibria result in very large chi-squares, of the order of $(N - m)/4m$ with Yates correction, so one or two such associations will dominate to total chi-square over the rest of the set.

As we shall now show, similar absolute constraints on the numerical possibilities for FET also exist for larger values of k.

k = 2–5: The results of applying Equations 1–3 for the cases of k = 2, 3, 4, and 5 are shown in Table 2. These suffice to make the points at issue. The last column of Table 2 gives approximate one-sided probabilities of observing different table fillings, at the lower and upper limits that m, the frequency of the rarer allele at the nonfocal site, may take. To understand the table, several points are in order.

Only values for $m \geq k$ are considered because the results for $m < k$ are already included in the table for smaller values of k. That is, the focal site is always the one with the more asymmetrical allele frequencies.

The values given in the last column of the table are numerical approximations for large samples where N is large enough compared with k that terms like $N/(N - k)$ are close to unity. Moreover, terms of smaller order in $1/N$ than those given have been ignored. Thus Table 2 is not meant to provide exact probabilities, but only to allow the test probabilities to be compared with the standard significance values of 0.05 and 0.01.

For each k the results have been separated into configurations and limits on m that produce coupling and those that produce repulsion disequilibria. These have different limits for m, which accounts for the appearance of irregularity in the table. So, for example, for $k = 4$ the (1,3) filling produces coupling disequilibria for $4 \leq m < N/4$, but repulsion equilibria in the range $N/4 < m < N/2$.

The results in Table 2 show some general features.

1. The possibility of observing a significant result is essentially independent of sample size, N, for repulsion tests. They are actually very weak functions of N. No repulsion disequilibria can be significant, even at the 5% level, for $k < 5$, unless m is close to $N/2$. To get a significant test at the 1% level requires that $k \geq 7$, and, again, there must be an intermediate polymorphism at the nonfocal site. To give a precise example, suppose $N = 100$, a very large sample for a sequence study, and let $k = 7$ and $m = 10$. In the data of SCHAEFFER and MILLER (1993), only 3.4% of all pairwise combinations have k and m as large or larger than these values. With these values even the most extreme repulsion filling (0,7) has a probability of 0.466 and it is not until m reaches a value of 34 that the (0,7) configuration has the conventional significance probability of 0.05!

   Figure 2 shows the combinations of k (vertical axis) and m (horizontal axis) that would be significant at the 0.05 (*) and 0.01 (**) levels, when the most extreme repulsion filling, $(0,k)$ is observed, assuming a sample size of $N = 100$. Our conclusion is that for most molecular data sets, there will be very few cases that allow a significant repulsion disequilibrium to be observed, even when the most extreme filling of the table occurs.
2. Coupling linkages have probabilities that are a decreasing function of sample size, *when the nonfocal site is very asymmetrical in allele frequency*, so that such linkages can be detected in large enough samples. The more extreme table fillings have probabilities that are inversely proportional to higher powers of N, so they will usually be significant even with moderate sample sizes. On the other hand, with more symmetrical frequencies at the nonfocal site, the coupling linkages have the same problems of detection as the repulsion linkages, because they are essentially independent of sample size.

Our general conclusion must be that linkage disequilibrium studies may not, in general, be a powerful tool in making inferences from sequence data because most allele frequencies are very asymmetrical. Analyses that look for excesses of significant D' values when $k < 5$ and these observed allele frequencies are far from 0.5:0.5, as they usually are in sequence studies, cannot produce significant results for repulsion disequilibria. Of course, the discovery of significant disequilibria, especially if these are repulsion disequilibria, and if they are clustered in a nonrandom pattern, would be powerful evidence of selection or other causes of recent common ancestry. Most of the pairwise comparisons, however, cannot give significant results and this lack of significant disequilibria is not very informative.

SIGN TESTS ON D

Even though the usual test for linkage disequilibrium is not useful for very asymmetrical allele frequencies, it might be possible to use such data to detect an overall excess of coupling or repulsion disequilibria by examining the sign of the disequilibrium. If there is no true linkage disequilibrium in the population from which samples are taken, then we can derive, from our previous results, the probability that observed sample values of D will be positive or negative, using the convention of Table 1 that a positive D means an excess of coupling

TABLE 2

***D'* and FET probabilities for small values of *k* with different table fillings**

k	$j,k-j$	*D'* sign	*D'* magnitude	Approximations for large *N* one-sided $F(j,k-j)$
2	2,0	+	1	$\frac{2}{N^2}\,(m=2)$ $0.25\left(m=\frac{N}{2}\right)$
	1,1	+	$\frac{n-m}{2N}$	$\frac{4}{N}\,(m=2)$ $0.75\left(m=\frac{N}{2}\right)$
	0,2	−	1	$\sim 1\,(m=2)$ $0.25\left(m=\frac{N}{2}\right)$
3	3,0	+	1	$\frac{6}{N^3}\,(m=3)$ $0.12\left(m=\frac{N}{2}\right)$
	2,1	+	$\frac{2n-m}{3n}$	$\frac{18}{N^2}\,(m=3)$ $0.50\left(m=\frac{N}{2}\right)$
	1,2	+	$\frac{n-2m}{3n}\left(m<\frac{N}{3}\right)$	$\frac{9}{N}\,(m=3)$ $0.70\left(m=\frac{N}{3}\right)$
	1,2	−	$\frac{n-2m}{3m}\left(m>\frac{N}{3}\right)$	$0.74\left(m=\frac{N}{3}\right)$ $0.50\left(m=\frac{N}{2}\right)$
	0,3	−	1	$\sim 1\,(m=3)$ $0.12\left(m=\frac{N}{2}\right)$
4	4,0	+	1	$\frac{24}{N^4}\,(m=4)$ $0.06\left(m=\frac{N}{2}\right)$
	3,1	+	$\frac{3n-m}{4n}$	$\frac{96}{N^3}\,(m=4)$ $0.31\left(m=\frac{N}{2}\right)$
	2,2	+	$\frac{2n-2m}{4n}$	$\frac{72}{N^2}\,(m=4)$ $0.69\left(m=\frac{N}{2}\right)$
	1,3	+	$\frac{n-3m}{4n}\left(m<\frac{N}{4}\right)$	$\frac{16}{N}\,(m=4)$ $0.68\left(m=\frac{N}{4}\right)$
	1,3	−	$\frac{n-3m}{4m}\left(m>\frac{N}{4}\right)$	$0.74\left(m=\frac{N}{4}\right)$ $0.31\left(m=\frac{N}{2}\right)$
	0,4	−	1	$\sim 1\,(m=4)$ $0.06\left(m=\frac{N}{2}\right)$
5	5,0	+	1	$\frac{120}{N^5}\,(m=5)$ $0.03\left(m=\frac{N}{2}\right)$
	4,1	+	$\frac{4n-m}{5n}$	$\frac{600}{N^4}\,(m=5)$ $0.19\left(m=\frac{N}{2}\right)$
	3,2	+	$\frac{3n-2m}{5n}$	$\frac{600}{N^3}\,(m=5)$ $0.50\left(m=\frac{N}{2}\right)$
	2,3	+	$\frac{2n-3m}{5n}\left(m<\frac{2N}{5}\right)$	$\frac{200}{N^2}\,(m=5)$ $0.66\left(m=\frac{2N}{5}\right)$
	1,4	+	$\frac{n-4m}{5n}\left(m<\frac{N}{5}\right)$	$\frac{25}{N}\,(m=5)$ $0.67\left(m=\frac{N}{5}\right)$
	2,3	−	$\frac{2n-3m}{5m}\left(m>\frac{2N}{5}\right)$	$0.68\left(m=\frac{2N}{5}\right)$ $0.50\left(m=\frac{N}{2}\right)$
	1,4	−	$\frac{n-4m}{5m}\left(m>\frac{N}{5}\right)$	$0.92\left(m=\frac{N}{5}\right)$ $0.19\left(m=\frac{N}{2}\right)$
	0,5	−	1	$\sim 1\,(m=5)$ $0.03\left(m=\frac{N}{2}\right)$

The values of $F(j, j-k)$ given are those for the largest and smallest allowable values of m, the absolute frequency of the rarer allele at the nonfocal site.

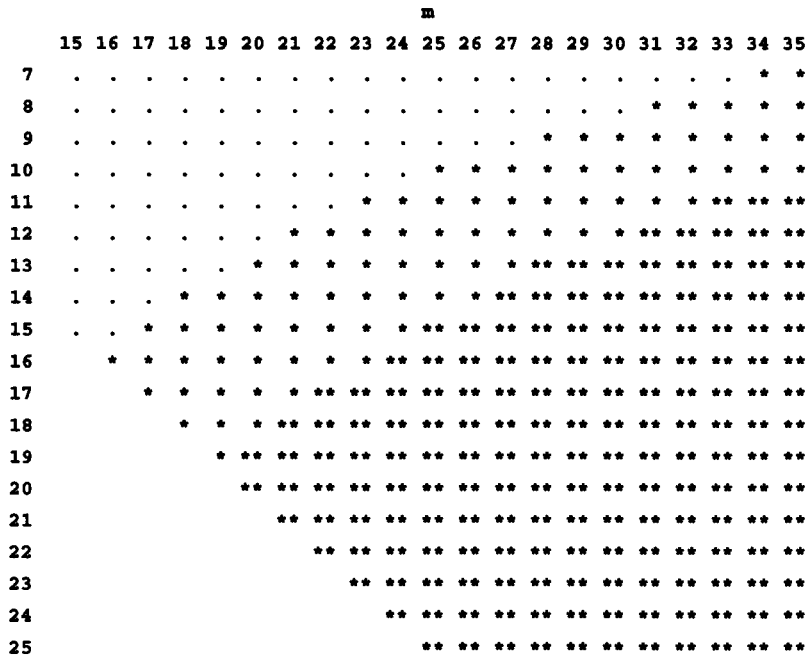

FIGURE 2.—Combinations of the absolute frequency of the rarer allele at the focal site (k) and the nonfocal site (m) that give the most extreme repulsion filling, (0,k) as not significant (·), significant at the 0.05 level (*) and 0.01 level (**). $N = 100$.

of rare alleles. If the observed proportions of positive and negative D values differ significantly from the expected proportion under the null hypothesis of site independence, then inference can be made about processes. If there are a very large number of site pairs to be classified as positive or negative, a simple test of proportions can be powerful even when the number of sequences examined is modest.

The procedure involves considering focal sites with each value of k and m separately. For a fixed value of k at the focal site, D will be negative only for certain table fillings and for certain bounds on m as shown in Table 2. For example, when $k = 4$, D will be negative for the filling (0,4) for all values of m and for those fillings (1,3) for which $m > N/4$. For $k = 5$, the conditions are more complicated, including all values of m for the filling (0,5) but only the values of $m > N/5$ and $m > 2N/5$ for the fillings (1,4) and (2,3), respectively. To determine, in general, the probability of a negative D for a given set of marginal frequencies m and k, the probabilities of various fillings of the table are calculated as usual from FET and the probabilities for those configurations that correspond to a negative D are summed. From Table 1, and bearing in mind that by definition m is the representation of the less frequent allele, the general rule for a configuration to give a negative D is simply

$$m > \frac{Nj}{k} \qquad (4)$$

Table 3 gives the probabilities of negative D values for various combinations of k and m for sample sizes $N = 25$, 50 and 100. The curious nonmonotonic behavior of the probabilities in the table arises because for a fixed k as m increases the probability of an extreme table filling becomes smaller and smaller, but at a critical value of m a new, less extreme, table filling now gives a negative D so that the probability jumps up again.

From Table 3, it can be determined what proportion of all pairwise site pairs of a given k and m should be negative, and a goodness-of-fit test between this expectation and then observed proportion of negative Ds can be run. The appropriate test statistic is the likelihood ratio statistic, G, which is a better approximation to the chi-square distribution than the more conventional chi square. Each k,m combination provides a goodness-of-fit G with 1 degree of freedom. In any real data set there will be only one or very few observations in most k,m classes, so the individual G values will not give reliable probabilities and nothing can be made of them separately. The sum of the G values, however, with the summed degrees of freedom will be quite close to the chi-square distribution under the null hypothesis, even for a moderate number of classes. This summed G is a test of whether, on the average, the number of positive and negative Ds conforms to expectation within each k,m class, so that a significant result could arise because there are too many positive Ds in some classes and too many negative Ds in others. To test for a general bias toward positive or negative associations the expected and observed numbers of positive and negative tests are summed over all k,m classes and a single goodness-of-fit test with one degree of freedom is carried out, using either chi-square or the likelihood ratio statistic, G.

Finally, the demonstration of a significant *statistical* deviation from random expectation should not be confused with a test of some particular *biological* hypothesis. Significant deviations mean only that there is some real

TABLE 3

Probability of negative D, depending on k, m, and sample size

	k				
m	1	2	3	4	5
1	0.900				
2	0.800	0.622			
3	0.700	0.467	0.292		
4	0.600	0.333	0.667	0.452	
5	0.500	0.222	0.500	0.262	0.500

$N = 10$.

	k											
m	1	2	3	4	5	6	7	8	9	10	11	12
1	0.960											
2	0.920	0.843										
3	0.880	0.770	0.670									
4	0.840	0.700	0.578	0.473								
5	0.800	0.633	0.496	0.383	0.292							
6	0.760	0.570	0.421	0.306	0.657	0.547						
7	0.720	0.510	0.355	0.693	0.564	0.443	0.337					
8	0.680	0.453	0.296	0.618	0.475	0.349	0.607	0.487				
9	0.640	0.400	0.713	0.542	0.391	0.637	0.501	0.374	0.593			
10	0.600	0.350	0.654	0.468	0.313	0.545	0.399	0.607	0.470	0.341		
11	0.560	0.303	0.593	0.396	0.622	0.452	0.649	0.496	0.352	0.534	0.393	
12	0.520	0.260	0.531	0.328	0.541	0.363	0.550	0.387	0.560	0.404	0.570	0.418

$N = 25$.

	k																							
m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0.980																								
2	0.960	0.921																							
3	0.940	0.882	0.827																						
4	0.920	0.845	0.774	0.709																					
5	0.900	0.808	0.724	0.647	0.577																				
6	0.880	0.772	0.676	0.589	0.513	0.444																			
7	0.860	0.737	0.630	0.536	0.454	0.384	0.323																		
8	0.840	0.703	0.586	0.486	0.401	0.330	0.690	0.622																	
9	0.820	0.669	0.544	0.440	0.354	0.707	0.630	0.555	0.483																
10	0.800	0.637	0.504	0.397	0.311	0.656	0.571	0.491	0.416	0.349															
11	0.780	0.605	0.466	0.357	0.699	0.604	0.513	0.430	0.355	0.618	0.544														
12	0.760	0.574	0.430	0.321	0.655	0.553	0.458	0.373	0.632	0.551	0.472	0.398													
13	0.740	0.544	0.396	0.725	0.611	0.503	0.406	0.659	0.571	0.485	0.404	0.624	0.547												

TABLE 3

Continued

	k																								
m	1	2	3	4	5																				
14	0.720	0.514	0.364	0.690	0.567	0.455	0.357	0.604	0.510	0.421	0.636	0.552	0.471	0.393											
15	0.700	0.486	0.334	0.654	0.524	0.409	0.652	0.549	0.451	0.362	0.570	0.482	0.398	0.589	0.507										
16	0.680	0.458	0.305	0.617	0.482	0.365	0.604	0.495	0.395	0.600	0.505	0.414	0.600	0.513	0.428	0.602									
17	0.660	0.431	0.736	0.580	0.440	0.674	0.554	0.442	0.642	0.539	0.440	0.622	0.529	0.438	0.608	0.520	0.434								
18	0.640	0.405	0.709	0.544	0.401	0.631	0.505	0.391	0.587	0.479	0.379	0.556	0.458	0.623	0.530	0.439	0.597	0.508							
19	0.620	0.380	0.680	0.507	0.362	0.588	0.457	0.649	0.532	0.421	0.595	0.490	0.390	0.551	0.454	0.606	0.513	0.421	0.569						
20	0.600	0.355	0.651	0.472	0.326	0.544	0.410	0.599	0.477	0.365	0.533	0.425	0.582	0.478	0.380	0.528	0.430	0.574	0.478	0.386					
21	0.580	0.331	0.621	0.436	0.654	0.501	0.365	0.549	0.423	0.589	0.471	0.624	0.514	0.407	0.552	0.449	0.588	0.488	0.390	0.525	0.427				
22	0.560	0.309	0.591	0.402	0.616	0.457	0.638	0.498	0.371	0.532	0.411	0.561	0.446	0.587	0.477	0.612	0.506	0.403	0.534	0.432	0.560	0.459			
23	0.540	0.287	0.561	0.368	0.578	0.415	0.593	0.448	0.606	0.474	0.620	0.497	0.380	0.517	0.404	0.535	0.425	0.552	0.445	0.569	0.464	0.586	0.482		
24	0.520	0.265	0.531	0.336	0.539	0.373	0.547	0.399	0.554	0.418	0.560	0.433	0.567	0.446	0.574	0.457	0.581	0.468	0.588	0.477	0.595	0.487	0.603	0.496	
25	0.500	0.245	0.500	0.305	0.500	0.334	0.500	0.351	0.500	0.363	0.500	0.371	0.500	0.377	0.500	0.381	0.500	0.384	0.500	0.387	0.500	0.388	0.500	0.389	0.500

$n = 50$.

	k																								
m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0.990																								
2	0.980	0.960																							
3	0.970	0.941	0.912																						
4	0.960	0.921	0.884	0.847																					
5	0.950	0.902	0.856	0.812	0.770																				
6	0.940	0.883	0.829	0.778	0.729	0.683																			
7	0.930	0.864	0.803	0.745	0.690	0.639	0.592																		
8	0.920	0.846	0.777	0.713	0.653	0.598	0.547	0.500																	
9	0.910	0.827	0.751	0.682	0.618	0.559	0.506	0.457	0.412																
10	0.900	0.809	0.727	0.652	0.584	0.522	0.467	0.417	0.371	0.330															
11	0.890	0.791	0.702	0.623	0.551	0.487	0.430	0.380	0.334	0.698	0.653														
12	0.880	0.773	0.679	0.595	0.520	0.455	0.397	0.345	0.706	0.657	0.608	0.561													
13	0.870	0.756	0.656	0.568	0.491	0.424	0.365	0.722	0.669	0.616	0.565	0.515	0.468												
14	0.860	0.738	0.633	0.542	0.463	0.394	0.336	0.689	0.632	0.576	0.523	0.472	0.423	0.378											
15	0.850	0.721	0.611	0.516	0.436	0.367	0.718	0.656	0.596	0.538	0.482	0.430	0.381	0.649	0.602										
16	0.840	0.704	0.589	0.492	0.410	0.341	0.689	0.624	0.560	0.500	0.443	0.390	0.656	0.605	0.555	0.507									
17	0.830	0.687	0.568	0.469	0.386	0.731	0.660	0.591	0.525	0.463	0.406	0.668	0.615	0.561	0.509	0.459	0.412								
18	0.820	0.671	0.548	0.446	0.362	0.706	0.631	0.559	0.491	0.428	0.371	0.630	0.574	0.518	0.465	0.414	0.636	0.588							
19	0.810	0.655	0.528	0.424	0.340	0.681	0.602	0.528	0.458	0.395	0.653	0.592	0.533	0.476	0.422	0.642	0.590	0.539	0.489						
20	0.800	0.638	0.508	0.403	0.319	0.655	0.574	0.497	0.427	0.363	0.618	0.555	0.494	0.436	0.382	0.598	0.544	0.492	0.440	0.392					
21	0.790	0.622	0.489	0.383	0.718	0.630	0.546	0.467	0.396	0.649	0.583	0.517	0.455	0.397	0.612	0.555	0.499	0.445	0.394	0.587	0.535				
22	0.780	0.607	0.470	0.364	0.697	0.605	0.518	0.438	0.367	0.617	0.548	0.481	0.418	0.631	0.571	0.512	0.455	0.401	0.591	0.538	0.485	0.434			
23	0.770	0.591	0.452	0.345	0.676	0.580	0.491	0.410	0.659	0.585	0.514	0.446	0.382	0.593	0.531	0.470	0.413	0.601	0.545	0.490	0.436	0.610	0.558		
24	0.760	0.576	0.435	0.327	0.654	0.555	0.464	0.383	0.630	0.553	0.480	0.411	0.620	0.555	0.491	0.429	0.616	0.557	0.499	0.443	0.614	0.560	0.506	0.453	
25	0.750	0.561	0.418	0.310	0.633	0.531	0.438	0.357	0.601	0.522	0.447	0.378	0.584	0.517	0.451	0.390	0.574	0.513	0.454	0.398	0.567	0.511	0.456	0.403	0.562

$N = 100$.

linkage disequilibrium, but this can arise from random processes, or selection or migration. To test whether the observed, statistically significant deviation could have arisen by chance from finite population processes requires an estimate of the *evolutionary stochastic* variance, using coalescent theory, rather than simply the sampling variation implicit in the goodness-of-fit test (see HUDSON 1985).

INDEPENDENT TESTS

To combine the tests from different pairs of loci into a single significance test requires that the pairwise tests be independent of each other.[1] It is common, in data exploration, to calculate all the D' values for all $n(n-1)/2$ pairs of n loci and to note which pairs are significant, in order to look for clusters of linked sites. While these tests are not statistically independent of each other, the dependence may not be very strong if the pairwise deviations from random association are not great. However, the dependence is absolute for the *signs* of nonindependent tests. If locus 1 and 2 are positively associated, and locus 2 and 3 are positively associated, then locus 1 and 3 must be positively associated. The criterion for independence of tests is that the graph that connects pairs of tested loci must have no closed loops. So pairs 1–2, 2–3, 3–4 are independent, but become dependent on the addition of 1–4. In general there will be $n-1$ independent pairs, and these can be constructed in a very large variety of ways. One simple scheme is to place the loci in any order and then take the pairs 1–2, 2–3, 3–4, . . . , $(n-1)$–n.

An obvious ordering is the actual order of sites in the sequence, but this does not necessarily provide the most powerful test for disequilibrium even though neighboring sites are expected to be in greatest disequilibrium. Inspection of Table 3 shows that k,m pairs in the upper left hand corner of the table, and pairs for even moderate m when $k = 1$, have very high probabilities of negative association. These combinations then have very low power to detect excess negative associations and provide powerful tests only against positive disequilibrium. Symmetrical power against both alternatives is achieved for those entries closest to 0.5, so pairwise schemes that are enriched for these entries will give more powerful tests. On the other hand there is always the danger that sorting data in this way will allow unconscious bias to prove a point, unless an a priori scheme is decided on.

ILLUSTRATIONS

RILEY *et al.* (1989) analyzed the *Xdh* region of *Drosophila pseudoobscura* by four-cutter analysis of 58 genomes, resulting in 78 polymorphic sites. The distribution of the allele frequencies is given above in Figure 1a. As discussed above, the frequency distribution is extremely J-shaped, with nearly every polymorphic site having its rare allele close to fixation. There are only two sites with nearly 50:50 polymorphisms. The following analysis excludes the three genomes that were sampled from the very distant Bogota population. To look for patterns of linked sites, all 3003 pairwise tests among the 78 sites can be examined. Of these, 1922 involve a singleton and 1773 of these latter associations are negative and therefore could not be significant for *any* sample size. Of the 219 positive associations for $k = 1$, 188 have an m greater than 2 and therefore cannot possibly be significant in the sample of 55 genomes. Thus, of all the pairwise tests possible, 65% could not possibly give significant results, for purely numerical reasons, just considering $k = 1$. This presents serious problems for detecting patterns of sites in disequilibrium, and clumps of linked sites will be apparently broken up by nonsignificant tests. Table 4 shows the proportion all the pairwise tests that were significant at the 0.05 level when various restrictions are placed on the minimum allowable k.

TABLE 4

Effect of minimum k on proportion of significant FETs in two data sets

	RILEY *et al.*		SCHAEFFER and MILLER	
Minimum k	No. of tests	Proportion significant	No. of tests	Proportion significant
1	3003	0.010	64261	0.032
2	1081	0.069	20301	0.073
5	231	0.078	5778	0.156
7	45	0.089	3741	0.213

To search for overall evidence of linkage disequilibrium, 77 pairwise comparisons among adjacent ordered sites can be made using the actual order of sites in the sequence. Figure 3 shows the distribution of probabilities from the 77 FETs and also the distribution of the these probabilities among the mere 23 tests left if singleton sites are excluded. Under the null hypothesis of no true disequilibrium, each probability interval on the abscissa should contain 20% of all tests. As the figure shows the actual test probabilities are strongly biased toward the very high probability classes and most of this skew remains even when the singletons are excluded. Of the 77 tests only 3, 3.8%, have probabilities less than the 0.05 significance level. Thus, FETs not only fail to give evidence of any overall linkage, but are strongly biased away from the expected probability distribution under the null hypothesis.

When we turn to tests based on the sign of D, there

[1] I am indebted to two anonymous reviewers who pointed out that I had failed to deal with this problem in an earlier version of the manuscript.

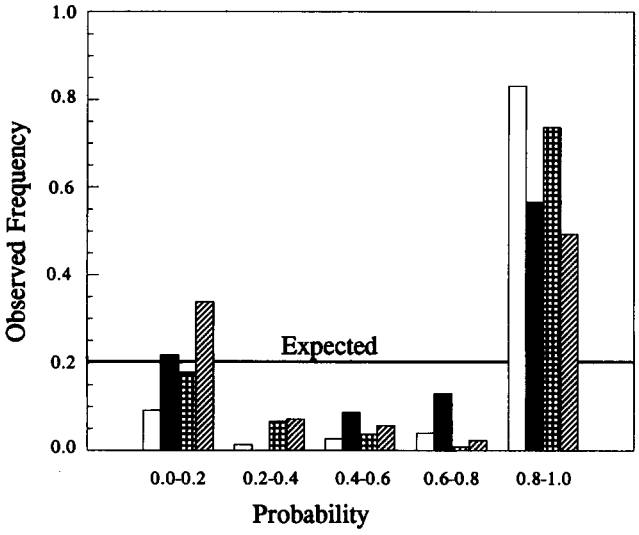

FIGURE 3.—Distribution of significance probabilities from Fisher's Exact Test. Data of RILEY *et al.*, total data (clear bars); excluding $k = 1$ (solid bars). Data of SCHAEFFER and MILLER, total data (cross-hatched bars); excluding $k = 1$ (shaded bars).

are again 77 pairwise tests among the adjacent sites on the sequence. There are 31 k,m classes from 1,1 to 8,9. The expected proportion of negative Ds for each class were computed exactly for $N = 55$, rather than interpolating in Table 3. The total G for goodness-of-fit is 39.21 with 31 degrees of freedom ($P = 0.14$), not corrected for continuity and 28.77 ($P = 0.58$) with the very conservative William's correction. If we exclude classes with $k = 1$, which have very high expected proportions of negative D's and therefore give very low power tests against an excess of these, there are 19 k,m classes with a total $G = 27.96$ ($P = 0.08$).The observed numbers of negative Ds is 68 and the expected number is 64.4 ($G = 1.34$, $P = 0.25$). Thus, RILEY *et al.*'s conclusion that is no clear evidence of any overall linkage disequilibrium is confirmed.

When the same analysis is done for the much larger experiment of SCHAEFFER and MILLER, a different result is seen. They sequenced the *Adh* region of *D. pseudoobscura* and found 359 polymorphic sites in 99 genomes. Table 4 shows that when $k = 1$ or 2 there is no evidence of any overall excess of significant pairs among the total set of 64261 comparisons, but for $k > 5$ there is a considerable excess of significant tests. To look for a significant overall linkage, 358 independent pairs were tested by FET with a distribution of probabilities shown in Figure 3. For these data too, the probabilities are strongly skewed the highest probability class when singletons are included, but when they are excluded there is now a significant excess of tests in the lowest probability class, giving evidence of real linkage disequilibrium (13% of tests have probabilities less than the 0.05 significance level).

Turning to tests based on the sign of D, there are 110 k,m classes among the 358 independent pair comparisons, ranging from 1,1 to 48,48. Using the expected proportions of negative tests from Table 3 ($N = 100$), the total G for goodness-of-fit is 177.05 with 110 degrees of freedom ($P < 0.001$) without correction for continuity and 138.74 ($P = 0.031$) with Williams correction. The observed number of negative D's is 286 and the expected number is 304.26 ($G = 6.72$, $P = 0.009$). Thus there is clear evidence of linkage disequilibrium in the SCHAEFFER and MILLER data, with an excess of coupling linkages.

IMPROVING POWER TO DETECT DISEQUILIBRIUM

The difficulty with the usual contingency test is not that it has low power, but that for purely numerical reasons it is not possible to achieve conventional significance levels when allele frequencies are low so that nearly all the polymorphisms in the sample are very asymmetrical.

The power of the sign test depends on the number of k,m classes, the number of pairwise comparisons in each class, and the distribution of pairs across various k,m classes. But these depend, in turn, not only on how much linkage disequilibrium is actually present, but also on the true distribution of allele frequencies in nature. A complete power analysis against various alternatives would then involve an exploration of a wide variety of a priori distributions of allele frequencies. It is possible, however, to draw some important conclusions about power without such an open-ended exploration.

Obviously, the power of the sign test can be increased by increasing either total number of sequences or the length of sequences. A choice between these depends in part upon whether added sequence length or added genome length can be regarded as being taken from the same universe as the less complete data set, but even if they are, the alternatives are not equally efficacious. Because k,m classes with $k = 1$ or 2 and low m values have very large expected proportions of negative Ds (Table 3), these classes have a high power to detect coupling disequilibria but very low power to detect repulsion disequilibria. Other k,m classes are equally powerful against both alternatives. Increasing the number of sequences has a different effect on the total number of k,m classes and the proportion of classes with high expected negative Ds than does increasing the sequence length. This can be shown using the RILEY *et al.* data and the SCHAEFFER and MILLER data as examples, because they show the J-shaped distribution of allele frequencies that is common in sequence samples. We can ask what would have been the effect on the number and distribution of k,m classes if the experimenters had sequenced only half as many genomes as opposed to the effect of having sequenced the same number of genomes, but for half the sequence length. A Monte

TABLE 5

Effect of reducing sample size and sequence length in two data sets

	N	Sites	Total pairs[a]	Total[a]	k, m classes $p(\text{neg}) > 0.85$[a]	G(total)	P
				RILEY *et al.*			
Original sample	55	78	77	31	123	39.21	0.148
Simulations (n = 40)	55	39	38	19.45 ± 0.371	10.73 ± 0.285	15.98	0.685
	27	78	40.53 ± 0.918	14.63 ± 0.286	11.63 ± 0.285	12.20	0.632
				SCHAEFFER and MILLER			
Original Sample	99	359	358	110	89	138.74	0.031
Simulations (n = 180)	99	179	178	66.96 ± 0.407	48.15 ± 0.416	75.24	0.232
	49	359	195.38 ± 1.27	60.83 ± 0.233	51.91 ± 0.236	83.04	0.028

[a] Values are means ± SE.

Carlo sampling simulation was carried out as follows. For half the sequence length, L, all $(L/2 + 1)$ successive windows of length $L/2$, beginning at the first polymorphic site, were laid down on the ordered sequence of polymorphic sites. This procedure does not quite simulate taking windows of a fixed length on the entire sequence since it holds the number of polymorphic sites constant, thus reducing the variance of the outcome, but it suffices to show the effect. For half the number of sequences, S, $(L/2 + 1)$ repeated random subsamples of size $S/2$ were taken with replacement from the original data of S sequences. The results of the simulations are shown in Table 5. The effect of increasing the number of genomes sampled as opposed to increasing the length of sequence is different for different aspects of the analysis. Both data sets show the same pattern and the differences discussed below are statistically significant at the 0.05 level or below as shown by the standard errors of the estimates, given in Table 5.

Decreasing the number of genomes sampled has a less drastic effect on reducing the total number of polymorphic sites and therefore, of the number of site pairs, than does decreasing the length of the sequence. This difference is quite small for the RILEY *et al.* data set, but very considerable for the larger SCHAEFFER and MILLER data set. So, on this ground power is increased more by increasing sequence length. The total number of k,m classes is decreased more by decreasing the sample size than by decreasing the sequence length. More k,m classes means more degrees of freedom for total goodness-of fit over all classes. On the other hand, the absolute number of classes with intermediate expected probabilities of negative associations is increased slightly more by increasing sequence length which means more power to detect negative associations but less power to detect positive associations. The difference in proportion of these classes, however, is considerable. Increasing the number of sequences has no statistically significant effect on the proportion of all k,m classes with intermediate probabilities of negative associations, but increasing the length of the sequence has a much larger effect, especially in the smaller data set of RILEY *et al.* The net result of changing the degrees of freedom while redistributing the frequencies of different k,m classes is not obvious. More degrees of freedom may give higher or lower rejection probabilities depending on how the total G changes. The last two columns of Table 5 show the net result. Reduction in the length of the sequence results in a greater increase in the P value than does reduction in the number of sequences sampled. This effect is small for the RILEY *et al.* data, but very considerable for the SCHAEFFER and MILLER data. Indeed, for these latter data, here is no observed loss of power at all when the number of sequences is cut in half! Our conclusion, then, is that for data of the kind seen in these two studies, more power is to be gained by increasing the length of sequence determined, than by increasing the number of genomes sampled. Again, it must be emphasized that this result, depends on whether the added sequence or can be taken from the same universe of linkage relations as the less complete data set.

The author is grateful to ANTONIO BARBADILLA, ANDREW BERRY, and SPENCER WELLS and to anonymous reviewers for very useful criticisms and constructive comments, to STEVE SCHAEFFER for providing a summary of data in a different form than published, and to RACHEL NASCA for setting up the production of the equations and tables. This work was supported by National Institute of General Medical Sciences grant GM-21179.

LITERATURE CITED

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

D'AGOSTINO, R. B., W. CHASE and A. BELANGER, 1988 The appropriateness of some common procedures for testing the equality of two independent binomial populations. Am. Statistician **42:** 198–202.

FU, Y. X., and J. ARNOLD, 1992 A table of exact sample sizes for use with Fisher's Exact Test for 2 × 2 tables. Biometric **48:** 1103–1112.

KENDALL, M., and A. STUART, 1979 *The Advanced Theory of Statistics,* Vol 2. Chas. Griffen, London.

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

OVERALL, J. E., H. M. RHOADES and R. R. STARBUCK, 1987 Small-sample tests for homogeneity of response probabilities in 2x2 contingency tables. Psych. Bull. **102:** 307–314

RILEY, M. A., M. E. HALLAS and R. C. LEWONTIN, 1989 Distinguishing the forces controlling genetic variation at the *Xdh* locus in *Drosophila pseudoobscura.* Genetics **123:** 359–369.

SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura.* Genetics **135:** 541–552.

TOCHER, K. D., 1950 Extension of the Neyman-Pearson theory of tests to discontinuous variates. Biometrika **37:** 130–144

UPTON, G. J. G., 1982 A comparison of alternative tests for the 2x2 comparative trial. J. R. Stat. Soc. Ser. A **45:** 86–105.

ZAPATA, C., and G. ALVAREZ, 1993 On the detection of non-random associations between DNA polymorphisms in natural populations of Drosophila. Mol. Biol. Evol. **10:** 823–841.

Communicating editor: A. G. CLARK