Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 9119–9123, December 1988
Genetics

Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci

(linkage disequilibrium/genetic epidemiology)

RANAJIT CHAKRABORTY* AND KENNETH M. WEISS†

*Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences at Houston, P.O. Box 20334, Houston, TX 77225; and †Department of Anthropology and Graduate Program in Genetics, The Pennsylvania State University, University Park, PA 16802

Communicated by William T. Sanders, May 25, 1988

ABSTRACT **Admixture between genetically different populations may produce gametic association between gene loci as a function of the genetic difference between parental populations and the admixture rate. This association decays as a function of time since admixture and the recombination rate between the loci. Admixture between genetically long-separated human populations has been frequent in the centuries since the age of exploration and colonization, resulting in numerous hybrid descendant populations today, as in the Americas. This represents a natural experiment for genetic epidemiology and anthropology, in which to use polymorphic marker loci (e.g., restriction fragment length polymorphisms) and disequilibrium to infer a genetic basis for traits of interest. In this paper we show that substantial disequilibrium remains today under widely applicable situations, which can be detected without requiring inordinately close linkage between trait and marker loci. Very disparate parental allele frequencies produce large disequilibrium, but the sample size needed to detect such levels of disequilibrium can be large due to the skewed haplotype frequency distribution in the admixed population. Such situations, however, provide power to differentiate between disequilibrium due just to population mixing from that due to physical linkage of loci—i.e., to help map the genetic locus of the trait. A gradient of admixture levels between the same parental populations may be used to test genetic models by relating admixture to disequilibrium levels.**

Admixture between two populations with different allele frequencies at two loci will produce a gametic association between these loci in any admixed population (1). Here, we refer to such gametic association as "mixture disequilibrium" to distinguish it from gametic association between closely linked loci. Such mixture disequilibrium will decay over time, but if the two loci are not linked, or their linkage is loose, nontrivial levels of disequilibrium may persist for long time periods. Even if recombination between loci occurs with a constant rate, mixture disequilibrium in an admixed population may remain over a substantial period of time, if the history of admixture is not very old. Chakraborty and Smouse (2) showed that in the presence of recombination, estimates of admixture from haplotype data may be error-prone, if the genetic assay of the admixed population is not done immediately after the admixture event. There are many instances where human populations have been formed through admixture of the same two stocks of racial groups, yet the degree of admixture varies among the admixed groups (with approximately the same historical depth of admixture). Such populations may offer an opportunity to use admixture as a tool for anthropological research.

Earlier, we (3) showed the utility of using admixed populations for fitting genetic models of inheritance of complex diseases. The objective of the present paper is to show that the observed levels of disequilibrium between any two loci in such an array of admixed populations may be used to detect their linkage relationship and to differentiate the case of mixture disequilibrium between loci from the disequilibrium that can be ascribed to genetic linkage.

MATHEMATICAL TREATMENT

Mixture Disequilibrium in an Admixed Population. As in the case of traditional admixture models, we consider two loci (A and B) that are not affected by selection. Let A and a, B and b be the two segregating alleles at these loci, respectively. Suppose that an admixed population (Z) obtains a fraction (m) of its genes from ancestral population X, and a fraction ($1 - m$) from ancestral population Y. We assume that the admixture event has taken place in a single pulse at generation 0, and the populations are surveyed t generations after this event. This theory is discussed in more detail elsewhere (ref. 2 and unpublished work). Let r denote the recombination rate between the A and B loci, and let $p_A(j)$, $p_a(j)$, $p_B(j)$, and $p_b(j)$ be the allele frequencies in population j ($j = X$, Y, or Z). Note that for any j, $p_a(j) = 1 - p_A(j)$ and $p_b(j) = 1 - p_B(j)$, and none of the allele frequencies change over generations, in the absence of selection and genetic drift (whose effects are ignored for the present discussion).

Under these assumptions, it is known (2, 4) that the mixture disequilibrium between the A and B loci in the admixed population Z, produced by admixture, at generation 0 can be written as

$$D_Z^{(0)} = mD_X^{(0)} + (1 - m)D_Y^{(0)} + m(1 - m)\delta_A\delta_B, \quad [1]$$

where $\delta_A = [p_A(X) - p_A(Y)]$ and $\delta_B = [p_B(X) - p_B(Y)]$. In addition, for any population j (X, Y, or Z), the mixture disequilibrium decays with time as a function

$$D_j^{(t)} = (1 - r)^t D_j^{(0)}. \quad [2]$$

Therefore, even if the parental populations (X and Y) are initially at linkage disequilibrium, if their allele frequencies are different ($\delta_A \neq 0$, $\delta_B \neq 0$), then the admixed population (Z) will exhibit mixture disequilibrium at the outset, due to its admixed origin. The disequilibrium will decay over time, due to recurrent recombination over generations, following the equation

$$D_j^{(t)} = (1 - r)^t m(1 - m)\delta_A\delta_B, \quad [3]$$

since in this case

$$D_Z^{(0)} = m(1 - m)\delta_A\delta_B. \quad [4]$$

Thus, for given values of δ_A and δ_B (allele frequency differences between the two parental populations) and mixture

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: lod, log-likelihood ratio.

proportion (m), we can determine what level of mixture disequilibrium will remain in an admixed population after t generations of the admixture event. This quantity is a function of the recombination rate r, which is a surrogate measure of the physical distance between the loci A and B. Eq. 3, therefore, shows that in an admixed population of known history of admixture (i.e., with known δ_A, δ_B, t, and m), the amount of mixture disequilibrium between two loci can be used to draw inferences regarding the linkage relationship between loci.

Fig. 1 shows the result of such computations, based on representative values of the allele frequencies in parental populations for two loci, A and B, by plotting the decay of mixture disequilibrium over time in the admixed population as a function of t for two values of r, for loci at different recombination distance from each other. The figure compares populations with very different allele frequencies (fixation at both loci) with populations with similar allele frequencies. The absolute value of D is highly dependent on these frequencies, and the range of D values is much larger if the gene frequencies are more disparate (note that the four panels have very different absolute ranges on the vertical scale). High admixture levels induce higher absolute D values because the range of possible D values is greater. The four panels of this figure cover ranges of t that encompass most situations in which human studies might be done (t up to 100 generations) and reasonable distances of marker from second locus in 10^3 base pairs (kb) if recombination is roughly 10^{-3} per kb.

It is clear from these computations that if in fact the two loci are linked ($r < ½$), even if the parental populations are at linkage equilibrium, admixed populations arising from populations of substantial allele frequency difference will exhibit mixture disequilibrium for a long period of time. Note that the absolute magnitude of disequilibrium in a population may not be very large at any point of time. This is expected because the magnitude of disequilibrium is also dependent on allele frequency (5), the admissible value of $D_Z^{(t)}$ being given by

$$-\min[p_A(Z)p_B(Z),\ p_a(Z)p_b(Z)] < D_Z^{(t)} < \min[p_A(Z)p_b(Z),\ p_a(Z)p_B(Z)]. \quad [5]$$

Power of Detection of a Given Level of Disequilibrium. Brown (6) considered the problem of determining the statistical power of detecting a given linkage disequilibrium, D_Z,

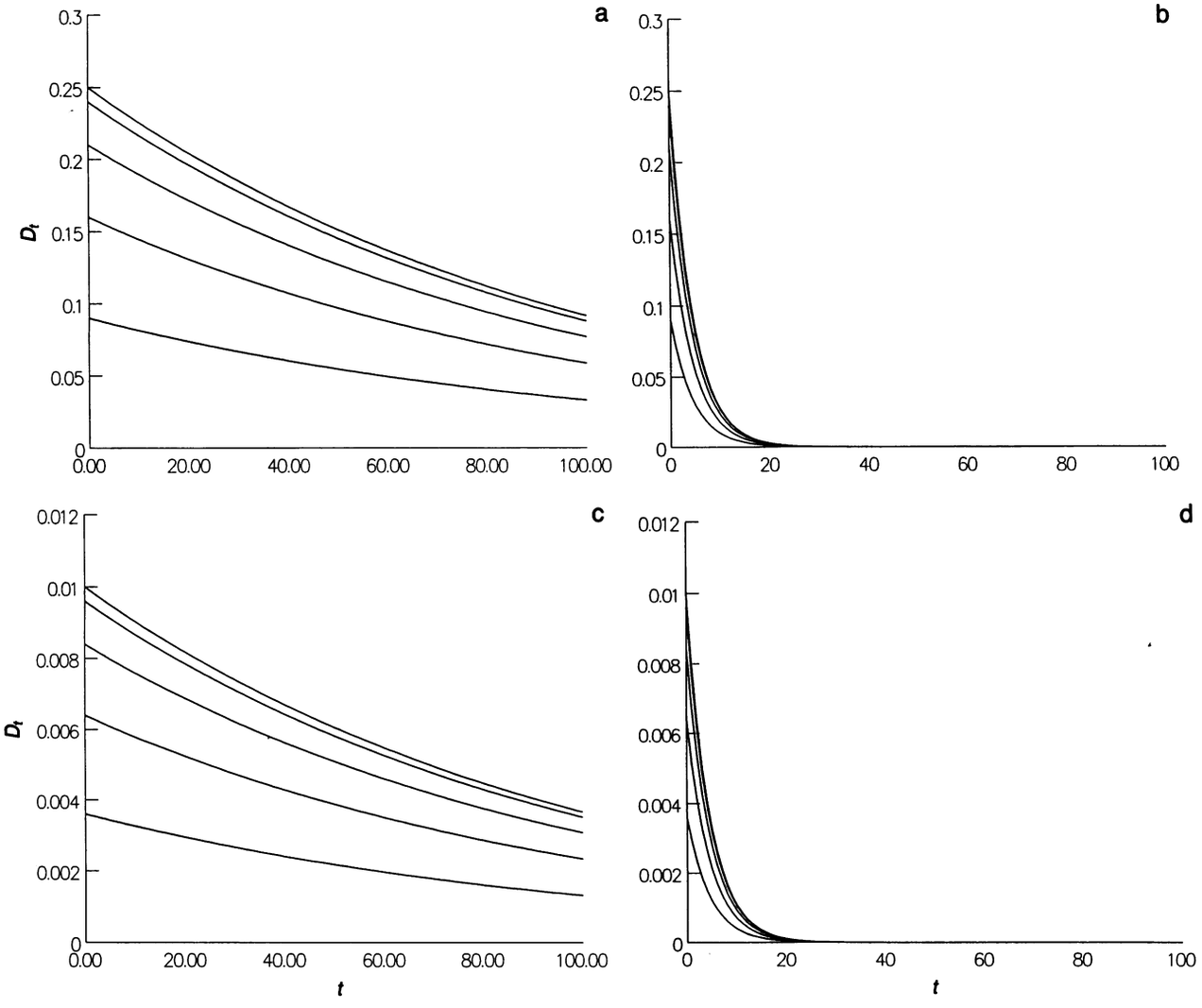

FIG. 1. Mixture disequilibrium values (D_t) as a function of time since admixture (t), for different levels of admixture (m). The four panels encompass some representative values of allele frequency differences of parental populations and recombination distance between loci A and B. Curves represent m values from 0.1 (bottom curve) to 0.5 (top curve) in increments of 0.1. (a) $p_A(X) = p_B(X) = 1.0$, $p_A(Y) = p_B(Y) = 0.0$, and $r = 0.01$. (b) $p_A(X) = p_B(X) = 1.0$, $p_A(Y) = p_B(Y) = 0.0$, and $r = 0.20$. (c) $p_A(X) = p_B(X) = 0.6$, $p_A(Y) = p_B(Y) = 0.4$, and $r = 0.01$. (d) $p_A(X) = p_B(X) = 0.6$, $p_A(Y) = p_B(Y) = 0.4$, and $r = 0.20$.

from a survey of n gametes. When the null hypothesis H_0: $D_Z^{(t)} = 0$ is tested against the alternative H_1: $D_Z^{(t)} \neq 0$, the test criterion for rejecting the null hypothesis H_0 is given by

$$\{C: |D_Z^{(t)}| > 1.96[p_A(Z)p_a(Z)p_B(Z)p_b(Z)]^{1/2}\} \quad [6]$$

for a 5% level of significance test (6). For a specific alternative $D_Z^{(t)} \neq 0$, the power of this test procedure is given by

$$\beta = 1 - [Q(r_2) - Q(r_1)], \quad [7]$$

where

$$r_1 = \frac{-1.96[p_A(Z)p_a(Z)p_B(Z)p_b(Z)]^{1/2} - D_Z(n)^{1/2}}{[p_A(Z)p_a(Z)p_B(Z)p_b(Z) + D_Z\varepsilon_A(Z)\varepsilon_B(Z) - D_Z]^{1/2}},$$

and

$$r_2 = \frac{1.96[p_A(Z)p_a(Z)p_B(Z)p_b(Z)]^{1/2} - D_Z(n)^{1/2}}{[p_A(Z)p_a(Z)p_B(Z)p_b(Z) + D_Z\varepsilon_A(Z)\varepsilon_B(Z) - D_Z]^{1/2}}.$$

$Q(r)$ is the cumulative (lower tail) probability of a standard normal variate r, and $\varepsilon_A(Z) = p_A(Z) - p_a(Z)$ and similarly for $\varepsilon_B(Z)$. This method of power evaluation, we may note, is more accurate than the approximation used by Brown (6), who used a one-sided normal integral to approximate Eq. 7. The accuracy of the normal deviate test (Eq. 6) is known to be better than that of the χ^2 test of allelic association (7). Fig. 2 shows the computations of the power, following Eq. 7, for the parameter values used in Fig. 1. Here the power is computed as functions of the standardized disequilibrium values, $D'_{(Z)} = D_{(Z)}/D_{(Z)}(\max)$, where $D_{(Z)}(\max)$ is the maximum absolute value, given by the bounds shown in expression 5 above.

Fig. 2 shows that for very disparate parental allele frequencies, even with relatively small sample sizes, the power to detect linkage disequilibrium is quite adequate so long as there is a reasonable amount of admixture. For intermediate parental allele frequencies the range of $D_{(Z)}$ is narrow, but even very small samples will be adequate to detect a small deviation from linkage equilibrium. This is so because in such instances the haplotype frequencies in the admixed population are more evenly distributed, and hence even in a small sample, all of them will be observed, making it easier to detect departure from equilibrium for small absolute levels of mixture disequilibrium. There is, however, a discontinuity at $p_A(X) = p_A(Y)$, at which point $D_{(Z)} = 0$; i.e., there is no mixture disequilibrium, even at the outset. It is interesting that in the former case (disparate parental frequencies), the power of the test can be small for very low levels of admixture, because of skewed haplotype frequencies in the admixed group.

Genetic Linkage Versus Mixture Disequilibrium Due to Mixture of Dissimilar Gene Pools. The mixture disequilibrium between two loci A and B in a population of admixed origin may be merely due to the admixture process alone and may not signify any linkage relationship between these loci. However, if we have an array of admixed populations, each of which arose from the same two parental populations, but which have undergone different levels of admixture, there will be a trend of disequilibrium values depending upon their admixture history. An approximate goodness-of-fit test for such a trend can be constructed based on the estimated linkage disequilibrium values in these populations, contrasting the disequilibrium values with their expectations based on Eq. 3. In a single admixed population of known history of admixture (i.e., with known parental allele frequencies and known values of t), a log-likelihood ratio test criterion may also be constructed to examine whether the observed value

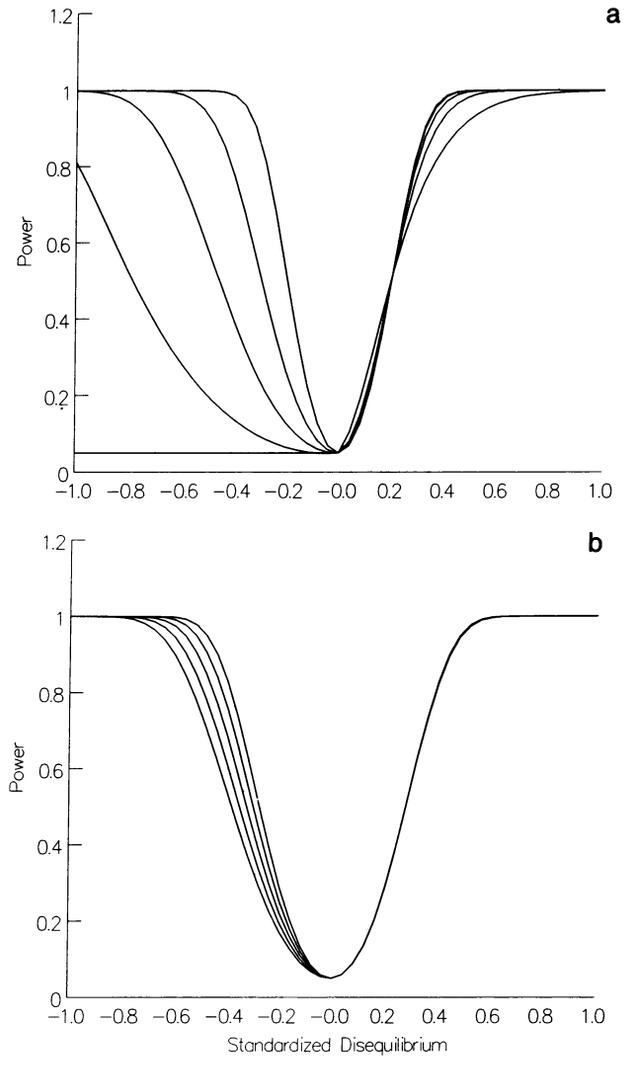

FIG. 2. Power to detect given levels of mixture disequilibrium as a function of admixture levels in an admixed population. The curves represent admixture (m), in increments of 0.1, ranging from $m = 0.1$ (bottom curve) to $m = 0.5$ (top curve). The two panels contrast the power as a function of disparity of allele frequencies in the parental populations. (a) $p_A(X) = p_B(X) = 1.0$, $p_A(Y) = p_B(Y) = 0.0$, and $n = 100$. (b) $p_A(X) = p_B(X) = 0.6$, $p_A(Y) = p_B(Y) = 0.4$, $n = 50$. Note that the disequilibrium is plotted as a fraction of its maximum and minimum, which are related to parental allele frequencies and admixture levels (see text).

of mixture disequilibrium is more likely to have come from the linkage of loci or from the process of admixture alone.

For this, let us consider an admixed population Z, as before, for which after t generations of the admixture event the four gametic frequencies in a survey of n gametes and their expected probabilities under the two hypotheses H_0: $r = \frac{1}{2}$, and H_1: a specific value of r, less than $\frac{1}{2}$, are shown in Table 1. Denoting these multinomial probabilities by $\pi_i(r)$ for H_1 and $\pi_i(\frac{1}{2})$ for the hypothesis H_0, for $i = 1, \ldots, 4$; we may write the log-likelihood ratio of these two hypotheses as

$$\text{lod} = \sum_{i=1}^{4} n_i[\log \pi_i(r) - \log \pi_i(\tfrac{1}{2})], \quad [8]$$

whose value can be computed for any $r < \frac{1}{2}$, given the other parameters in a particular survey. In analogy with linkage tests from family data, the value of the lod score itself can be used to decide whether H_0 or H_1 is in conformity with the observed data. Since the specific value of the lod score de-

Table 1. Observed number of two locus gametic types and their probabilities under the hypotheses of linkage and admixture

Gamete	Observed frequency	Probability under the hypothesis	
		H_1: $r < ½$	H_0: $r = ½$
AB	n_1	$p_A p_B + D_{H_1}$	$p_A p_B + D_{H_0}$
Ab	n_2	$p_A p_b - D_{H_1}$	$p_A p_b - D_{H_0}$
aB	n_3	$p_a p_B - D_{H_1}$	$p_a p_B - D_{H_0}$
ab	n_4	$p_a p_b + D_{H_1}$	$p_a p_b + D_{H_0}$

$D_{H_1} = (1 - r)^t m(1 - m)\delta_A\delta_B$; $D_{H_0} = m(1 - m)\delta_A\delta_B/2^t$.

pends upon the observed gametic frequencies (n_i values), it it is not possible to answer in advance the strategic question whether or not discrimination between these two hypotheses is possible from a particular survey design. However, a conservative decision may be reached if we want to evaluate the *expected* lod score for any given r in a survey with sample size n. For this we replace the n_i values of Eq. **8** by their respective expectations, $E(n_i|r) = n\pi_i(r)$, to obtain the expected lod score as a function of r, given by

$$E_r(\text{lod}) = n\sum_{i=1}^{4} \pi_i(r)[\log \pi_i(r) - \log \pi_i(½)], \qquad [9]$$

which can be plotted against values of r ($0 < r \leq ½$) for any given sample size n. Fig. 3 shows some representative values of such computations based on some parameter values used in our earlier computations (Figs. 1 and 2).

Fig. 3 shows that with disparate parental allele frequencies, the power to resolve the two hypotheses is very high under realistic sampling circumstances, for reasonably closely linked loci, because in such an instance the decay of initial mixture disequilibrium will be small due to the lack of recombinants due to close genetic linkage. With highly overlapping parental allele frequencies, even with much larger samples and small recombination rate, genetic linkage cannot be demonstrated statistically, because even with linkage actual recombinant haplotypes will be indistinguishable from existing parental haplotypes.

When the expected lod score is large enough to suggest true linkage (i.e., maximum lod score ≥ 3, in the tradition of usual convention in genetic epidemiology), then it will be profitable to collect detailed family data from the population. In that case such data will be informative for the segregation and mapping of the trait locus.

DISCUSSION

The above theory suggests an interesting way to make use of the past history of human populations as a research strategy in anthropological and epidemiological studies. During the evolution of various racial groups of humans, genetic isolation over long time periods, documentable in historical, linguistic, archeological, and other kinds of data, has allowed mutations and gene substitutions to generate genetic diversity with a highly patterned geographic distribution. Some of these mutations have caused complex phenotypes to evolve with their specific geographic distributions. While the mode of inheritance of such complex phenotypes may be perplexing, their geographic distribution often suggests involvement of genetic factors. Examples of such complex phenotypes are disease susceptibilities: diabetes in Amerindians and Polynesians, rheumatoid arthritis in Amerindians, skin cancer in Caucasians, and hypertension in Blacks.

In plant and animal genetics, when strains carrying unusual phenotypes are discovered, the inheritance of such traits is often studied by careful crossing of these strains with those that do not carry them. In human genetics, this is not possible. However, during recent centuries large-scale movement of populations over continental distances has given rise to admixed populations, where very different gene pools have mixed. Such admixed populations are reminiscent of genetic crosses.

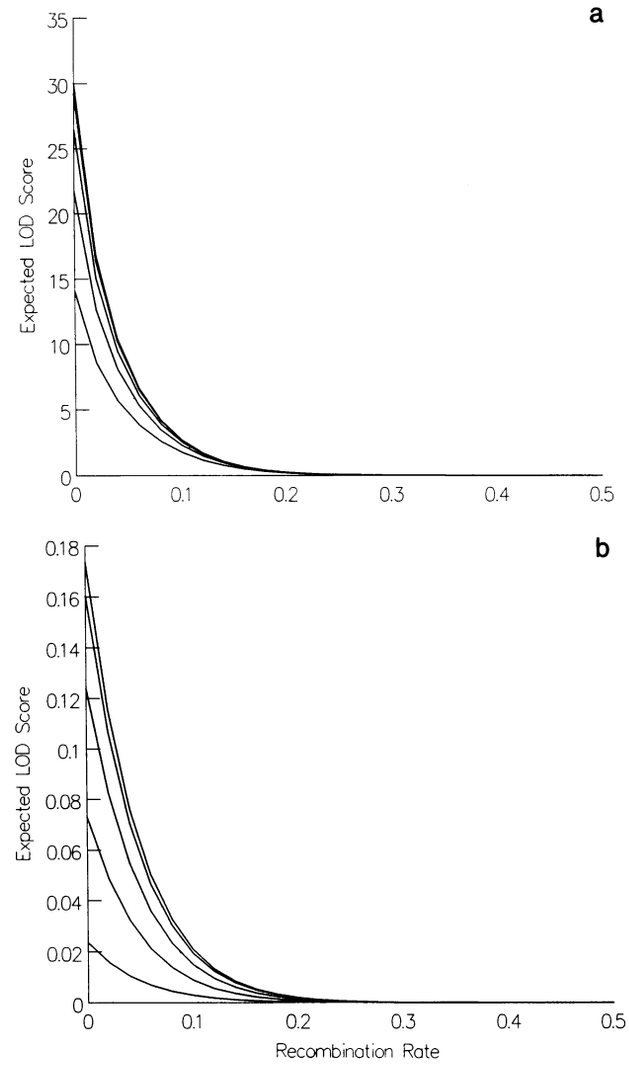

FIG. 3. Expected lod scores to differentiate between disequilibrium due to admixture and that due to genetic linkage, as a function of recombination rate between loci, for different levels of admixture. The curves represent admixture (m) in increments of 0.1, ranging from $m = 0.1$ (bottom curve) to $m = 0.5$ (top curve). The two panels show the pattern for different parental allele frequencies and sample size (n = number of haplotypes sampled). (*a*) $p_A(X) = p_B(X) = 1.0$, $p_A(Y) = p_B(Y) = 0.0$, $t = 10$, and $n = 100$. (*b*) $p_A(X) = p_B(X) = 0.6$, $p_A(Y) = p_B(Y) = 0.4$, $t = 10$, and $n = 500$.

Before discussing the advantages of genetic studies in admixed groups, we must mention the consequences of the simplified assumptions made in our model. We assumed that the admixed group is formed by a single "pulse" of admixture. In nature, however, admixture is a continual process that occurs over many generations. We will deal with this problem elsewhere (unpublished work); when admixture continues for a certain number of generations, larger mixture disequilibrium will be exhibited in the admixed group, and it will persist for a longer period of time after the admixture process ceases. Though qualitatively similar, continuous admixture reduces the power of discrimination between mixture disequilibrium and physical linkage.

The admixing of groups with very different gene frequencies, especially if important alleles are nearly fixed in one and nearly absent in another of the groups, will lead to a high

level of segregating matings. Such gene variants may be present in greatly different (perhaps completely different) haplotype backgrounds in the two parental populations, making linkage studies from randomly sampled families quite feasible in admixed populations.

In many of these circumstances, not only is genetic variation great between the parental populations, but the time since admixture is short (a few centuries, or on the order of $t = 10$ generations). Marker loci (restriction fragment length polymorphisms) within a small recombination distance (say, $r < 0.1$) of the trait locus will have good statistical properties in regard to drawing genetic inference as well as in mapping genes. It is also often possible to ascertain samples from a variety of populations with varying levels of admixture between the same two parental populations. Elsewhere it has been shown how admixture may be used to infer the existence of genetic etiological factors in complex phenotypes such as non-insulin-dependent diabetes mellitus in Amerindians (8–10), as well as how models of genetic causation may be tested with data representing a gradient of population admixture levels (3).

We thank Drs. P. E. Smouse and J. C. Long for their comments on the manuscript. This work was supported in part by Grants GM20293 and CA19311 from the National Institutes of Health and a grant from the Wenner–Gren Foundation for Anthropological Research.

1. Nei, M. & Li, W.-H. (1973) *Genetics* **75,** 213–219.
2. Chakraborty, R. & Smouse, P. E. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 3071–3074.
3. Chakraborty, R. & Weiss, K. M. (1986) *Am. J. Phys. Anthropol.* **70,** 489–503.
4. Thomson, G. & Klitz, W. (1987) *Genetics* **116,** 623–632.
5. Lewontin, R. C. (1964) *Genetics* **49,** 49–67.
6. Brown, A. H. D. (1975) *Theor. Popul. Biol.* **8,** 184–201.
7. Chakraborty, R. (1984) *Genetics* **108,** 719–731.
8. Chakraborty, R. (1986) *Ybk. Phys. Anthropol.* **29,** 1–44.
9. Chakraborty, R., Ferrell, R. E., Stern, M. P., Haffner, S. M., Hazuda, H. P. & Rosenthal, M. (1986) *Genet. Epidemiol.* **3,** 435–454.
10. Hanis, C. L., Chakraborty, R., Ferrell, R. E. & Schull, W. J. (1986) *Am. J. Phys. Anthropol.* **70,** 433–441.