GENOMICS **29,** 311–322 (1995)

A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping

B. DEVLIN*,[1] AND NEIL RISCH*,†,[2]

*Departments of Epidemiology and Public Health and †Genetics, Yale University School of Medicine, New Haven, Connecticut

Received March 22, 1995; accepted June 30, 1995

Linkage mapping generally localizes disease genes to 1- to 2-cM regions of chromosomes. In theory, further refinement of location can be achieved by population-based studies of linkage disequilibrium between disease locus alleles and alleles at adjacent markers. One approach to localization, dubbed simple disequilibrium mapping, is to determine the relative location of the disease locus by plotting disequilibrium values against marker locations. We investigate the simple mapping properties of five disequilibrium measures, the correlation coefficient Δ, Lewontin's D', the robust formulation of the population attributable risk δ, Yule's Q, and Kaplan and Weir's proportional difference d under the assumption of initial complete disequilibrium between disease and marker loci. The studies indicate that δ is a superior measure for fine mapping because it is directly related to the recombination fraction between the disease and the marker loci, and it is invariant when disease haplotypes are sampled at a rate higher than their population frequencies, as in a case-control study. D' yields results comparable to those of δ in many realistic settings. Of the remaining three measures, Q, Δ, and d, Q yields the best results. From simulations of short-term evolution, all measures show some sensitivity to marker allele frequencies; however, as predicted by analytic results, Q, Δ, and d exhibit the greatest sensitivity to variation in marker allele frequencies across loci.

INTRODUCTION

Linkage or pedigree analysis remains the fundamental paradigm by which genetic epidemiologists map loci contributing to inherited disorders (Ott, 1991). In fact, numerous genes having a major effect on human diseases have been mapped to within 1 cM using such analyses. Further refinement in location using family studies is difficult because recombinations are rarely observed even within the large pedigrees that would be required for finer mapping of these loci (Boehnke, 1994).

Consequently, it will often be the case that linkage mapping of disease loci leaves about 1 Mb of DNA to be searched by the molecular geneticist, which can be a daunting amount unless there are natural candidate genes in the region (e.g., Shiang *et al.,* 1994). Any method that narrows the amount of DNA to be searched would be important. One such method uses linkage disequilibrium to refine the location of the disease locus. Conditional on disease status, the linkage disequilibrium between a mutant allele at a disease locus and other alleles at flanking markers is complete (sensu Clegg *et al.,* 1976) at the instant the mutation occurs. When evolutionary forces can be ignored, including marker and disease locus mutation, any decay in disequilibrium is due solely to recombination. Under this ideal scenario, and provided that the time since the disease mutation is not too long, the pattern or curve of disequilibrium between disease and marker loci will exhibit a single maximum that occurs at the disease locus. Consequently, the amount of linkage disequilibrium between a disease allele and closely linked genetic markers may yield valuable information regarding the location of the disease gene.

We term this method of linkage disequilibrium mapping simple disequilibrium mapping because it uses only the pattern of pairwise disequilibrium values across loci to infer the approximate location of the disease locus. It is the method most commonly applied, although it is clear that other methods of disequilibrium mapping may make more efficient use of the data. For instance, Hill and Weir (1994) advance a maximum likelihood method for disequilibrium between two loci, a disease locus and marker locus, assuming that the population itself is in a steady state of constant population size and selective pressures (or neutrality). When these assumptions are met, their method will have some very desirable properties for localizing disease genes. Hästbacka *et al.* (1992) suggest another method of fine mapping using linkage disequilibrium, which is formulated specifically for recently founded popula-

[1] To whom correspondence should be addressed at current address: Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. Telephone: (412) 268-8874. Fax: (412) 268-7828.

[2] Current address: Department of Genetics, Stanford University School of Medicine, Stanford, CA.

0888-7543/95 $12.00

tions. Again, this method depends on certain assumptions about the evolutionary process, specifically exponential growth and a single disease-producing chromosome in the founding population, as well as knowledge of when the mutation first occurred. For a refinement of this method, see Kaplan *et al.* (1995). Regardless of the competing methods, simple disequilibrium mapping is a valid descriptive tool that molecular biologists frequently find useful for fine mapping.

Indeed, the problem of refined mapping of a disease locus via linkage disequilibrium is not just of theoretical interest. It has proved valuable in some notable instances. In the most celebrated case, the cystic fibrosis gene was mapped using a combination of molecular and population genetic techniques, including linkage disequilibrium mapping (Kerem *et al.*, 1989; Rommens *et al.*, 1989; Riordan *et al.*, 1989). Ozelius *et al.* (1992a,b) and Risch *et al.* (1991, 1995) have recently narrowed the location of the torsion dystonia gene to a small region of chromosome 9 (9q34) using linkage disequilibrium mapping in the Ashkenazi Jewish population. Linkage disequilibrium mapping has also been employed to localize the gene for Friedreich Ataxia using French Canadian, Italian, and Louisiana Acadian populations (Fujita *et al.*, 1990; Hanauer *et al.*, 1990; Richter *et al.*, 1990; Pandolfo *et al.*, 1990; Sirugo *et al.*, 1992), myotonic dystrophy using Caucasian populations (Harley *et al.*, 1991; Tsilfidis *et al.*, 1991), Lubag's disease using a Philippine population (Graeber *et al.*, 1992; Wilhelmsen *et al.*, 1992), diastrophic dysplasia (Hästbacka *et al.*, 1992, 1994), and infantile neuronal ceroid lipofuscinosis (Hellsten *et al.*, 1993) using a Finnish population, Huntington disease using Caucasian populations (Huntington Disease Collaborative Research Group, 1993), Wilson disease using various populations, including Caucasians (Petrukhin *et al.*, 1993; Bowcock *et al.*, 1994), and polycystic kidney disease using a Scottish population (Snarey *et al.*, 1994). For marker loci, Jorde *et al.* (1994) found that linkage disequilibrium was an excellent predictor of physical distance in the adenomatous polyposis coli region of chromosome 5 using a Caucasian population (see also Daiger *et al.*, 1989; Jorde *et al.*, 1993).

There are, however, reasons to be cautious about the use of linkage disequilibrium for fine mapping. Weir (1989) and Hill and Weir (1994) have been pessimistic about this technique because linkage disequilibrium is influenced by other phenomena besides recombination, namely mutation, drift, breeding system, and selection (Nei, 1987). These population genetic phenomena can mask the impact of recombination, leading at the least to a large variance in the disequilibrium values among loci (Weir, 1989; Hill and Weir, 1994). At worst, it could result in no relationship or even a misleading relationship between physical distance and linkage disequilibrium (Litt and Jorde, 1986; Thompson *et al.*, 1988; Walter and Cox, 1991).

In addition, recombinant mapping or linkage analysis is fundamentally different from simple disequilibrium mapping. Recombinant mapping places specific bounds on the location of the disease gene, whereas simple disequilibrium mapping can indicate only the likely location of the gene. The precision of this likely location depends on evolutionary phenomena, as well as the locations of the marker loci relative to the disease locus (detailed below).

Clearly, if simple disequilibrium mapping is to be useful, optimal strategies must be employed. One feature of the analysis that has not received much attention is the measure of disequilibrium. Numerous measures of linkage disequilibrium have been devised over the past 60 years of population genetic research, none of which has been shown to be optimal for simple disequilibrium mapping. Various measures have been used, and when two measures were compared (Jorde *et al.*, 1994), the conclusion was that they differed very little.

In this report, we discuss the fine-mapping properties of five commonly used measures of linkage disequilibrium. We first elaborate the relationships between these measures of disequilibrium and their relationships to other standard statistical quantities. We then show, via simple deterministic examples, analytic methods, and stochastic simulations, that the choice of linkage disequilibrium measure can have a substantial impact on the accuracy and interpretability of the simple disequilibrium mapping method. In what follows we restrict our discussion to marker loci having two alleles and a disease locus having two alleles, a "disease" and a "normal" allele. Thus the haplotypes for the disease locus and any single marker locus can be arrayed in a 2 × 2 table. Even if the marker has more than two alleles, the association is usually with only one (e.g., under complete disequilibrium), so marker alleles can be classified into two classes. The assumption of a single mutation at the disease locus is a far more important assumption.

MEASURES OF LINKAGE DISEQUILIBRIUM

Hedrick (1987) has reviewed the numerous measures of linkage disequilibrium. In his review, Hedrick demonstrates the conditions under which the measures, or at least a subset thereof, are highly correlated.

Consider two loci, each locus having two alleles: a disease allele and a normal allele segregate at the first locus, and two marker alleles segregate at the other locus. The layout and notation of the 2 × 2 table from a sample from the population are given in Table 1.

In Table 1, n_{11} is the number of haplotypes in the sample carrying the disease allele and marker allele A1, n_{1+} is the number of haplotypes bearing the A1 allele, n_{+1} is the number of haplotypes bearing the disease allele, and n is the total number of haplotypes sampled. Dividing these quantities by n yields the frequencies and marginal probabilities (denoted by p) from the sample (Table 2).

Conditional probabilities are written similarly to the

TABLE 1

Layout and Notation for Sample Haplotype Frequencies in a 2 × 2 Table

Marker	Disease allele	Normal allele	
A1	n_{11}	n_{12}	n_{1+}
A2	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

unconditional probabilities in Table 2. For instance, the probability of having allele A1 in the haplotype, given that the disease allele is present, is denoted $p_{1|+1} = p_{11}/p_{+1}$. Likewise, the probability of having the normal allele in the haplotype, given that the marker allele is A2, is given by $p_{2|2+} = p_{22}/p_{2+}$.

Naturally the p's are only sample estimates of some underlying unknown parameters, denoted by π's. We use π's in the definitions that follow, with the understanding that these unknown quantities are estimated from the observed sample quantities.

The basic component of many measures of disequilibrium is the difference between the observed and the expected (under independence) number of haplotypes bearing the disease allele and the A1 allele or its equivalent expressions:

$$\begin{aligned} D &= \pi_{11} - \pi_{1+}\pi_{+1} = \pi_{22} - \pi_{2+}\pi_{+2} \\ &= -\pi_{12} + \pi_{1+}\pi_{+2} = -\pi_{21} + \pi_{2+}\pi_{+1} \\ &= \pi_{11}\pi_{22} - \pi_{12}\pi_{21}. \end{aligned}$$

According to Hill and Weir (1994), the most frequently used measure of disequilibrium is the square of the standardized measure

$$\Delta = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})^{1/2}}$$

or Δ^2. Δ is commonly squared to remove the arbitrary sign introduced when the marker alleles are labeled.

Another common measure, introduced by Lewontin (1964), is defined as

$$D' = \begin{cases} \dfrac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\min(\pi_{1+}\pi_{+2},\ \pi_{+1}\pi_{2+})} & D > 0 \\[2ex] \dfrac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\min(\pi_{1+}\pi_{+1},\ \pi_{+2}\pi_{2+})} & D < 0 \end{cases}.$$

The quantity in the denominator is the absolute maximum D that could be achieved given the table margins.

These measures are related to standard statistical measures of association. In particular, Δ is the correlation coefficient for a 2 × 2 table (Hill and Robertson, 1968). Δ is also proportional to Haberman's (1973) adjusted residuals for the 2 × 2 table

$$r_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{(\hat{m}_{ij}(1 - p_{i+})(1 - p_{+j}))^{1/2}},$$

where m_{ij} is the expected number in cell ij.

Another association measure that finds frequent use in epidemiology and has also been used to study linkage disequilibrium in Levin's (1953) population attributable risk δ^*. This quantity is defined as

$$\begin{aligned} \delta^* &= \frac{\pi_{1+}}{\pi_{+1}}(\pi_{1|1+} - \pi_{1|2+}) \\ &= \frac{\pi_{1+}(\phi - 1)}{1 + \pi_{1+}(\phi - 1)}, \end{aligned}$$

where $\phi = \{\pi_{11}/\pi_{1+}\}/\{\pi_{21}/\pi_{2+}\}$, the relative risk. An approximation for this measure of association or disequilibrium was first used in the population genetics context by Bengtsson and Thomson (1981) (see also Thomson, 1981). Specifically, by appealing to the odds ratio approximation to the relative risk (e.g., Breslow and Day, 1980), one obtains after some algebra an approximation for the population attributable risk that is robust to sampling disease haplotypes at a higher rate than their population frequencies (i.e., case-control sampling)

$$\delta = \frac{\pi_{1|+1} - \pi_{1|+2}}{\pi_{2|+2}} = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{+1}\pi_{22}}$$

(Levin and Bertell, 1978). Subsequent to Bengtsson and Thompson's (1981) research on HLA associations, δ has been used by Ozelius *et al.* (1992a,b) and Risch *et al.* (1991, 1995) for simple disequilibrium mapping. Most recently it has been rederived and used for disequilibrium mapping by Lehesjoki *et al.* (1993), who referred to it as P_{excess}, and by Terwilliger (1995), who referred to it as λ; however, these measures are simply δ.

The measure δ^* is not entirely new to population genetics. In fact, when the disease is rare and haplotypes are sampled at random, $\delta \doteq \delta^* = D'$:

TABLE 2

Notation for Estimated Haplotype, Marker Allele, and Disease Allele Frequencies in a 2 × 2 Table

Marker	Disease allele	Normal allele	
A1	p_{11}	p_{12}	p_{1+}
A2	p_{21}	p_{22}	p_{2+}
	p_{+1}	p_{+2}	1

$$\delta^* = \frac{\pi_{1+}}{\pi_{+1}}(\pi_{1|1+} - \pi_{1|2+})$$
$$= \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{+1}\pi_{2+}},$$

after some algebra. But

$$D' = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\min(\pi_{1+}\pi_{+2}, \pi_{+1}\pi_{2+})}$$
$$= \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{+1}\pi_{2+}}$$

when the table is structured so that D is positive and the disease is rare relative to the associated marker allele frequency (see also Thomson, 1981). The denominator of D' is typically $\pi_{+1}\pi_{2+}$ for rare diseases and random sampling because it is the minimum if $\pi_{12} - \pi_{21} = \pi_{1+} - \pi_{+1} > 0$. This condition is met whenever the associated marker allele is more common in the population than the disease allele. Notice that $\delta^* = D'$ and δ differ only in their denominators, $\pi_{+1}\pi_{2+}$ versus $\pi_{+1}\pi_{22}$. For a rare disease, π_{21} is small, so $\pi_{2+} = \pi_{21} + \pi_{22} \doteq \pi_{22}$. However, $D' \neq \delta$ under case-control sampling.

Another epidemiologic measure (Nei and Li, 1980), which was specifically recommended for disequilibrium mapping when case-control sampling is employed (Kaplan and Weir, 1992) is the difference in proportions d

$$d = \frac{\pi_{11}}{\pi_{+1}} - \frac{\pi_{12}}{\pi_{+2}} = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{+1}\pi_{+2}}.$$

Other natural epidemiologic measures, again robust to case-control sampling, have found some use in population genetics, specifically the odds ratio λ and Yule's (1900) Q (e.g., Clegg *et al.*, 1976; Nei and Li, 1980; Olson and Wijsman, 1994). Recall that

$$\lambda = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

and therefore ranges from zero to infinity, while

$$Q = \frac{\lambda - 1}{\lambda + 1} = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}}$$

ranges between negative one and one. The last expression for Q shows its relationship to δ. In fact, the numerators of Δ, D', δ, d, and Q are all equal to D, and these measures differ only in their denominators, which serve to standardize D (Table 3).

TABLE 3

Disequilibrium Measures Commonly Used for Fine-Scale Mapping

Symbol	Formula
Δ	$\frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})^{1/2}}$
D'	$\frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{+1}\pi_{2+}}$
δ	$\frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{+1}\pi_{22}}$
d	$\frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{+1}\pi_{+2}}$
Q	$\frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}}$

Note. The notation, with π's substituted for p's, is defined in Table 1. Note that the numerators of the measures are identical, but the denominators, which standardize the measures, are not. The formula for D' is a special case discussed in the text.

In what follows, we focus on five measures of disequilibrium (or association): Δ, D', δ, d, and Q. One might conjecture that they all yield equivalent information for simple disequilibrium mapping. However, we illustrate by some deterministic examples, analytic results, and evolutionary simulations that this is not the case. In our examples we assume that the two haplotypes for each individual can be determined (as, for example, for a recessive disease or for multiplex families with a dominant disease). However, our conclusions also apply to the more general situation.

THE PERFORMANCE OF LINKAGE DISEQUILIBRIUM MEASURES FOR SIMPLE FINE MAPPING

Deterministic Calculations

Predicted patterns in populations. Imagine there are 50 founders of a new population. One individual carries a dominant disease allele D at a locus of interest. On the chromosome bearing the disease locus, let there be three biallelic markers on each side of the disease locus, and one biallelic marker at the disease locus itself (M_0). The two markers adjacent to the disease locus are equidistant from it (denote them M_1 and $M_{1'}$), the next furthest pair are also equidistant from the disease locus (M_2 and $M_{2'}$, keeping the "primes" on the same side), and likewise for the furthest markers (M_3 and $M_{3'}$). Because the pairs are equidistant from the disease locus, the recombination rates between disease locus and marker are assumed to be equal; define them to be $\theta_1 = \theta_{1'} = 0.002$, $\theta_2 = \theta_{2'} = 0.007$, and $\theta_3 = \theta_{3'} = 0.012$. In the population, let the allele frequency vectors for these seven markers, from M_3 to $M_{3'}$, be (0.25, 0.75), (0.5, 0.5), (0.25, 0.75), (0.5, 0.5), (0.5, 0.5), (0.25, 0.75), (0.5, 0.5) and let the first value in each tuple correspond to the frequency of the marker allele carried on the chromosome bearing the disease allele.

Assume that the allele frequencies of the markers and the disease locus remain relatively stable from generation to generation. However, after the initial appearance of the disease allele, recombination erodes the disequilibrium between the disease allele and the marker alleles. At founding, or generation $t = 0$, the joint frequency of the disease allele and the "leading" marker allele (i.e., the first allele at each locus) is 0.01. When we sample this population at $t = 100$ generations, the expected frequencies of disease chromosomes bearing the original marker alleles are, from M_3 to $M_{3'}$, 0.0047, 0.0075, 0.0086, 0.01, 0.0091, 0.0062, and 0.0065.

Notice that the pairs of equidistant markers do not have equal joint frequencies of disease allele and marker alleles even though their recombination rates are the same and, initially, the joint frequencies are equal. This is because, at a given locus, recombination need not generate a new haplotype. The rate at which recombination generates new haplotypes depends on marker allele frequencies, which are not equal for the equidistant markers even in the founding generation. Consider a marker locus M_i at generation $t = 0$. The joint frequency of disease allele D and marker allele 1 is given by $\pi_{11}^{(t=0)} = \pi_{1+}\pi_{+1} + D_{t=0}$. ($D$ is the disequilibrium measure defined above.) The disequilibrium for the 100th generation is $D_{t=100} = (1 - \theta_i)^{100} D_{t=0}$. The joint frequency at $t = 100$ is then $\pi_{11}^{(t=100)} = \pi_{1+}\pi_{+1} + D_{t=100}$. All joint frequencies for our example can be calculated in this fashion.

These differences between frequencies, for both marker alleles and joint frequencies of disease and marker alleles, are important for fine mapping. As Hedrick (1987) emphasized, measures of disequilibrium such as Δ can be difficult to interpret when loci differ in their allele frequencies. Other measures, such as D' and δ, are more easily interpreted. Furthermore, the ability to determine correctly the location of the disease locus from the pattern of disequilibrium values depends on the measure used. For example, consider the disequilibrium values from our population at generation $t = 100$ (Fig. 1, top). The maximum for Δ and d are not at the disease locus, but at an adjacent marker. (To make the measures comparable in Fig. 1, Δ and d have been rescaled by multiplying the set of values for each measure and scenario by a constant.) In addition, these disequilibrium measures yield multimodal patterns of disequilibrium. By contrast, δ and D' exhibit almost identical behavior: they are unimodal and essentially symmetric and their maximum is at the disease locus. Finally, Q has a maximum at the appropriate location, but it shows marked deviation from symmetry.

If we examine the population in an early generation, say ($t = 5$), the results would be even more dramatic (Fig. 1, bottom). In this case, it would be difficult to define even a region to search for the disease locus if the researcher uses Δ or d as the measure of disequilibrium. D' and δ place the disease locus in the appropriate location, although the peak itself is little differentiated from other locations. Q shows behavior similar to that of D' and δ in that its maximum is the same; however, Q has another peak at the extreme left marker and it also has other asymmetries such as the "right shoulder." From these examples, it should be clear that the choice of disequilibrium measure can be important for simple disequilibrium mapping.

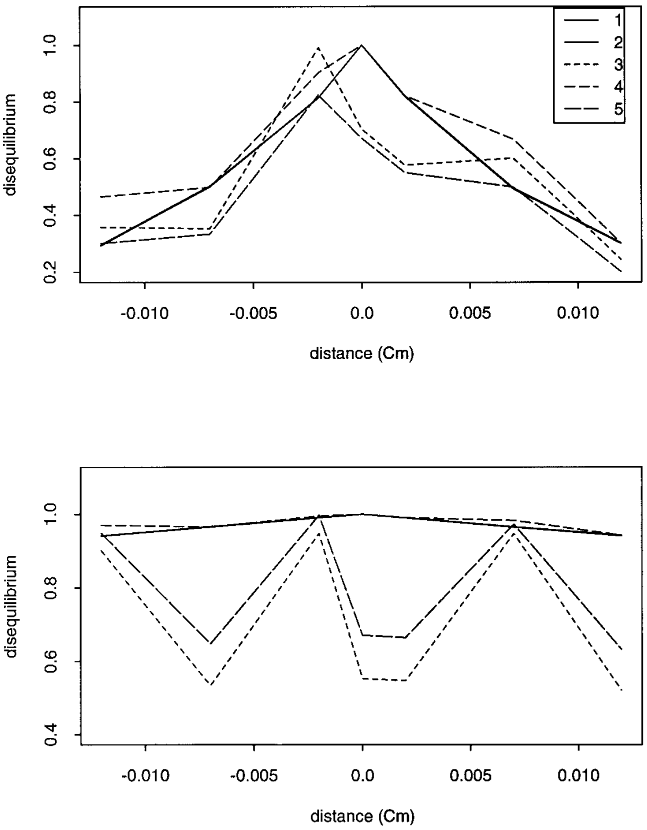

FIG. 1. Linkage disequilibrium versus recombination fraction for five disequilibrium measures: 1 = D', 2 = δ, 3 = Δ, 4 = Q, and 5 = d. The patterns displayed are generated by a model population (see text for details). Both D' and δ display an ideal pattern (overlapping solid lines) for simple disequilibrium mapping.

To see why this is so, recall the expression for disequilibrium at generation n, $D_n = (1 - \theta)^n D_0$. Ideally we desire a measure that is a function of θ only, for instance

$$(1 - \theta)^n = \frac{D_n}{D_0} = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{+1}\pi_{22}}.$$

Our rationale for the denominator of this expression is as follows. Under the assumption of initial complete linkage disequilibrium and no change of disease allele frequency over time, $\pi_{11} + \pi_{21} = \pi_{+1}$ is the best estimate in generation n of the initial disease allele frequency and hence π_{11}; π_{21} at generation 0 is 0. Thus $\pi_{+1}\pi_{22}$ is the best estimate of D_0, the initial amount of linkage disequilibrium. Reexpressing the measures of linkage disequilibrium in terms of $(1 - \theta)^n$ is revealing. As shown above, $(1 - \theta)^n$ is exactly δ.

For D', when $D > 0$ and $\pi_{1+} > \pi_{+1}$,

$$(1 - \theta)^n = D'\left(1 + \frac{\pi_{21}}{\pi_{22}}\right).$$

Thus, the relationship between D' and θ depends somewhat on haplotype and marker allele frequencies. For rare diseases, however, $\pi_{21} \ll \pi_{22}$ usually, making D' essentially only a function of θ. An exception occurs when one of the marginal marker allele frequencies is rare: the effect of rare π_{2+} can be seen in the expression above; the effect of rare π_{1+} is to change the denominator of D'. In fact, when $\pi_{1+} < \pi_{+1}$,

$$(1 - \theta)^n = D'\left(\frac{(1 + \pi_{12}/\pi_{11})(1 + \pi_{21}/\pi_{22})}{(1 + \pi_{21}/\pi_{11})}\right),$$

making the relationship between D' and θ in this exceptional case dependent on haplotype and marker allele frequencies.

The relationship between Q and θ,

$$(1 - \theta)^n = Q\left(1 + \frac{\pi_{12}/\pi_{22} - 1}{\pi_{11}/\pi_{21} + 1}\right),$$

also reveals a dependence on haplotype frequencies. The coefficient of Q potentially ranges between $(0, \infty)$, although extreme values occur only when the unassociated marker allele frequency is small.

The relationship between d and θ can be deduced from the relationship between δ and θ because $d = \pi_{22}\delta/\pi_{+2}$ and therefore

$$(1 - \theta)^n = \frac{d\pi_{+2}}{\pi_{22}}.$$

Thus, d depends on haplotype and marker allele frequencies.

Finally, it is apparent that the relationship between Δ and θ is obscured by marginal allele frequencies:

$$(1 - \theta)^n = \Delta\,\frac{1}{\pi_{22}}\sqrt{\frac{\pi_{1+}}{\pi_{+1}}(1 - \pi_{1+})(1 - \pi_{+1})}.$$

When all five measures are compared, it becomes apparent that δ is the measure most directly related to θ. Furthermore, reexpressing δ in terms of haplotype frequencies,

$$\delta = \frac{\pi_{11}/\pi_{21} - \pi_{12}/\pi_{22}}{\pi_{11}/\pi_{21} + 1},$$

shows that it depends only on the relative frequencies π_{11}/π_{21} and π_{12}/π_{22} and not on the marginal marker allele frequencies (see also Thomson, 1981). Thus it is the ideal measure for simple disequilibrium mapping.

TABLE 4

Expected Disequilibrium after 100 Generations of Random Recombination between Marker and Disease Loci for Five Measures of Disequilibrium, as a Function of Associated Marker Allele Frequency and Recombination Fraction θ

		Marker allele frequency				
Measure	θ	0.167	0.333	0.5	0.667	0.833
δ	0.003	0.742	0.742	0.742	0.742	0.742
	0.006	0.550	0.550	0.550	0.550	0.550
	0.009	0.407	0.407	0.407	0.407	0.407
D'	0.003	0.740	0.740	0.740	0.740	0.740
	0.006	0.548	0.548	0.548	0.548	0.548
	0.009	0.405	0.405	0.405	0.405	0.405
Q	0.003	0.900	0.815	0.744	0.684	0.634
	0.006	0.790	0.650	0.521	0.479	0.424
	0.009	0.677	0.600	0.408	0.340	0.292
Δ	0.003	0.166	0.105	0.074	0.053	0.033
	0.006	0.123	0.078	0.055	0.039	0.025
	0.009	0.091	0.058	0.041	0.029	0.018
d	0.003	0.623	0.499	0.374	0.249	0.125
	0.006	0.461	0.367	0.277	0.184	0.092
	0.009	0.340	0.273	0.205	0.136	0.068

Note. The disease and marker loci were initially in complete disequilibrium, with a disease allele frequency of 0.01.

Other measures, such as D', are also proportional to θ, at least under certain circumstances. To illustrate the behavior of the five measures, we first use deterministic calculations similar to our first example. All calculations are based on complete disequilibrium between a disease allele (occurring with frequency 0.01) and marker allele at generation 0, which breaks down over 100 generations by recombination alone. The results (Table 4) at generation 100 reveal the sensitivity of Δ, Q, and d to haplotype and marker allele frequency variation. Low-frequency alleles far from the disease locus give higher values of Δ, d, and Q than closer, high-frequency alleles. δ and D' are insensitive to such variation.

Thus, from deterministic calculations, it appears that either δ or D' is ideal for simple disequilibrium mapping. However, the attributes of D' depend strongly on its denominator, which, in turn, depends on marker allele and disease allele frequencies. D' is usually directly related to θ; however, for common disease alleles, or for case-control sampling, D' need not be directly related to θ.

Effect of case-control sampling. A common strategy in linkage disequilibrium studies is to sample higher proportions of diseased individuals relative to their population frequencies, as in a case-control study. If the study, by design, samples disease chromosomes with probability π^*_{+1} as compared to π_{+1}, the value $c =$

TABLE 5

Haplotype, Marker, and Disease Frequencies for *c*-fold Increased Sampling of Disease Haplotypes

Marker	Disease allele	Normal allele	
A1	$c\pi_{11}$	$\pi_{1\|+2}(1 - c\pi_{+1})$	$\dfrac{cD + \pi_{12}}{\pi_{+2}}$
A2	$c\pi_{21}$	$\pi_{2\|+2}(1 - c\pi_{+1})$	$\dfrac{\pi_{22} - cD}{\pi_{+2}}$
	$c\pi_{+1}$	$1 - c\pi_{+1}$	1

π^*_{+1}/π_{+1} is a convenient means of expressing the effect of such case-control sampling on the relative frequencies in a 2 × 2 table (Table 5). When we discuss case-control sampling, we take $c = 50$ and $\pi_{+1} = 0.01$. This sampling yields equal numbers of disease and normal chromosomes, analogous to the typical strategy for case-control studies.

Whereas haplotype frequencies change, δ and Q are unaffected by case-control sampling. This invariance follows formally from the fact that both measures are functions of the odds ratio, which is invariant to case-control sampling (Edwards, 1963). Likewise, d is also unaffected by case-control sampling, as can be seen by substituting the adjusted haplotype frequencies (Table 5) into the equations for d.

The other two measures are affected by case-control sampling in some way. We examined the effect of case-control sampling on the pattern of disequilibrium across marker loci by supposing that a grid of markers surround the disease locus and the marker allele frequencies vary systematically between 0.083 and 0.917. Other attributes are identical to those used to develop Table 4, except that $c = 50$ (Table 5), so that equal numbers of disease and normal haplotypes are observed.

For D', the pattern is frequently multimodal and, for small (<0.1) marker frequencies, the maximum disequilibrium need not occur at the proximate marker locus (data not shown). These results are quite different from our results for random sampling, in which the pattern was always unimodal with a maximum at the proximate locus. The impact of case-control sampling on D' is mediated, in large part, through the choice of denominator. (Recall that the relationship between D' and θ depends critically on which of two terms is the minimum.) For case-control sampling, the denominator of D' is the

$$\min\left\{c\pi_{+1}\,\frac{\pi_{22} - cD}{\pi_{+2}},\,(1 - c\pi_{+1})\,\frac{cD + \pi_{12}}{\pi_{+2}}\right\},$$

and the first expression is the minimum if $c\pi_{12} - \pi_{1|+2}(1 - c\pi_{1+}) > 0$. This expression, which parallels that found for random sampling, can be greater than zero only when the sampled disease haplotype frequency is less than the associated marker allele frequency.

The results using Δ as the measure of disequilibrium differ markedly from those using D' (data not shown). In this instance, the multimodality of the pattern changes very little from that obtained from the population, although case-control sampling leads to an increase in number of times the proximate marker locus shows maximum disequilibrium. With some algebra, it can be shown that this increase occurs because case-control sampling changes the relationship between disequilibrium values at different loci relative to the values obtained from random sampling.

Impact of Stochastic Factors

We examined the impact of evolutionary forces by simulation of short-term population evolution. Details of the simulations are given in the Appendix. In brief, each population initially consisted of 2000 chromosomes (i.e., 1000 individuals), which then grew over 100 generations to a size of 100,000 chromosomes. Population expansion occurred at a constant exponential rate. Recombination occurred at random, as did reproduction. No mutation occurred.

To examine systematically the impact of variation in marker allele frequencies, we simulated populations of chromosomes having three marker loci, at distances θ = 0.001, 0.004, 0.007 from the disease locus. Initial marker allele frequencies were either of three values, 0.1, 0.5, 0.9, and all possible combinations of those values for different loci were examined (i.e., 27 sets). For each combination of marker allele frequencies, 80 populations were simulated. The initial disease allele frequency was set to 0.01; if the frequency of this allele dropped below 0.005 during any generation, the simulation was reinitialized at generation zero. Marker allele frequencies were allowed to go to zero (rarely occurred), in which case the locus was ignored as it was not polymorphic.

Two types of data were examined from each population: the disequilibrium pattern for the population as a whole (population pattern) and the disequilibrium pattern for case-control sampling (with $c = 50$, in expectation); specifically, 200 disease chromosomes and 200 normal chromosomes were sampled. We recorded, for each set of allele frequencies, the fraction of the time the nearest marker exhibited the greatest disequilibrium and the mean square error (MSE), computed as the sum of the squared recombinational distance between the disease locus and the marker exhibiting maximum disequilibrium between it and the disease locus. Ideally the MSE would be $(0.001)^2 = 1E - 6$, which would occur if the nearest locus always exhibited maximum disequilibrium. MSE is an appropriate measure of variability in this instance because it naturally incorporates both variance and any bias into a single statistic.

The simulation results agree with the deterministic calculations in terms of the average performance over all sets of allele frequencies. For the population pattern, the nearest marker locus exhibited the greatest disequilibrium the highest fraction of times with both δ and D' (83.5%), followed by Q (81.2%), then d (52.9%), and finally Δ (48.3%). MSE shows the identical pattern, with both δ and D' having the smallest MSE (4.94E-6), followed by Q (5.17E-6), then d (1.38E-5), and finally Δ (1.55E-5). Taking the square root of the MSE, we have 0.0022, 0.0023, 0.0037, and 0.0039, respectively. (Recall that MSE emphasizes occasional larger deviations relative to a measure such as the average absolute deviation.)

For case-control sampling, δ outperforms all other measures in terms of the pattern of maximum disequilibrium and MSE (81.2% and 5.39E-6), followed by D' (78.1% and 6.55E-6), Q (76.6% and 6.79E-6), Δ (55.1% and 1.29E-5), and d (52.8% and 1.39E-5). Taking the square root of the MSE yields 0.0023, 0.0026, 0.0026, 0.0036, and 0.0037, respectively. Note that, as predicted by the deterministic calculations, the performance of δ and D' are now distinct and that the performance of Δ improves with case-control sampling, relative to the population patterns. Moreover, because we sampled a relatively large number of haplotypes (200 disease, 200 normal), the impact of sampling error per se is small. Naturally smaller sample sizes will increase the MSE of simple disequilibrium mapping.

Substantive patterns are hidden by these data summaries. As demonstrated by the deterministic calculations, the poor performance of Δ and d is due to a bias involving the magnitude of allele frequencies. Large disequilibrium values are associated with small allele frequencies and vice versa; thus, both measures, when used for simple disequilibrium mapping, frequently cause it to be an inconsistent estimator of the marker nearest the disease locus (i.e., the estimator does not converge to the true answer as the sample size tends to infinity). However, this bias could also fortuitously work in the investigator's favor. For instance, when the nearest marker's associated allele frequency is small, and other associated marker allele frequencies are much larger, the proximate marker will almost invariably show a large disequilibrium value using either Δ or d. For the simulations, when associated allele frequencies for furthest to nearest markers were initially set to 0.5, 0.9, 0.1, the largest disequilibrium value occurred at the proximate marker 100% of the time (population level). Alternatively, for the configuration 0.9, 0.1, 0.9, the largest disequilibrium value never occurred at the proximate marker for either measure. The bias is illustrated in Table 6, which presents the results for case-control sampling only.

Q shows behavior similar to that of Δ and d. As the deterministic calculations suggest, however, it is less sensitive to the magnitude of marker allele frequencies (Table 6). The behavior of δ and D' in the stochastic simulations deviate somewhat from the deterministic calculations. The deterministic calculations suggest that both measures should be unaffected by the magnitude of marker allele frequencies, whereas the simulations clearly show that the performance of these measures for simple disequilibrium mapping also changes with the configuration of allele frequencies (Table 6). These measures are most affected when the frequency of an associated marker allele is large. For instance, when the associated allele frequency configuration was 0.9, 0.9, 0.1 for furthest to nearest markers, the largest disequilibrium value occurred for the proximate marker only 62.5% of the time (population patterns) and the MSE was 9.1E-6. Conversely, when the associated allele frequency configuration was 0.1, 0.1, 0.1, the largest disequilibrium value occurred for the proximate marker 96.25% of the time (population pattern) and the MSE was 2.0E-6.

Most of this behavior is attributable to the variance in δ and D'. Because these measures are essentially identical under many circumstances, we discuss only δ. The asymptotic standard error for $\log(1 - \delta)$ is

$$\left(\frac{\pi_{11}}{n_{+1}\pi_{21}} + \frac{\pi_{12}}{n_{+2}\pi_{22}}\right)^{1/2}$$

(Walter, 1975), and therefore the asymptotic standard error of δ increases as the unassociated marker allele frequency, π_{2+}, tends toward zero. While we are less interested in statistical sampling than in genetic sampling (sensu Weir, 1990), genetic sampling can be thought of as repeated statistical sampling. Therefore, the sensitivity of δ to the unassociated allele frequency, as revealed by the formula for its asymptotic standard error variance formula, is pertinent.

In another set of simulations, we allowed initial marker allele frequencies to vary at random between the limits 0.15 and 0.85 and specified seven marker loci with recombination, relative to the disease locus of 0.009, 0.006, 0.003, 0, 0.003, 0.006, 0.009. Other simulation conditions were the same as those described previously, except that 200 populations were simulated. In this case, the largest disequilibrium value for δ, D', and Q always occurred at the disease locus for both the population and the case-control sampling scenarios. For Δ, the largest disequilibrium value occurred with the proximate marker 44% of the time for the population and 54.5% of the time for case-control sampling. For d, the largest disequilibrium value occurred with the proximate marker 47% of the time for the population and 48% of the time for case-control sampling.

DISCUSSION

At the instant a new "disease" mutation occurs, the disease allele is associated with alleles at other polymorphic loci in the region. In particular, the disease locus is in complete linkage disequilibrium (Clegg *et al.*, 1976) with other loci in the region. When it is reasonable to assume that the disease locus was initially

TABLE 6

Simulation Results of Short-Term Evolution and Subsequent Case-Control Sampling by Initial Marker Allele Frequency (Furthest to Nearest Marker from Left to Right)

			Disequilibrium measures				
Allele frequency			D'	δ	Δ	Q	d
0.1	0.1	0.1	0.90 (3.33)	0.96 (1.98)	0.95 (1.75)	0.95 (2.16)	0.95 (1.75)
0.1	0.1	0.5	0.68 (9.63)	0.95 (1.79)	0.18 (15.91)	0.55 (9.85)	0.19 (14.88)
0.1	0.1	0.9	0.78 (8.38)	0.85 (3.71)	0.06 (20.56)	0.64 (9.84)	0.06 (18.91)
0.1	0.5	0.1	0.98 (3.63)	0.98 (1.41)	1.0 (1.00)	0.99 (1.77)	1.0 (1.00)
0.1	0.5	0.5	0.83 (4.84)	0.91 (2.37)	0.44 (26.40)	0.73 (12.61)	0.36 (29.17)
0.1	0.5	0.9	0.94 (3.82)	0.95 (1.81)	0.05 (40.65)	0.84 (7.50)	0.04 (40.83)
0.1	0.9	0.1	0.68 (7.22)	0.68 (5.93)	0.99 (2.06)	0.78 (4.55)	0.99 (2.09)
0.1	0.9	0.5	0.60 (10.48)	0.66 (6.63)	0.51 (24.12)	0.61 (10.63)	0.45 (27.52)
0.1	0.9	0.9	0.81 (5.31)	0.85 (3.84)	0.05 (46.05)	0.75 (8.71)	0.03 (47.75)
0.5	0.1	0.1	0.83 (4.60)	0.90 (5.08)	0.93 (3.69)	0.91 (2.94)	0.95 (2.75)
0.5	0.1	0.5	0.85 (3.34)	0.96 (1.62)	0.21 (12.95)	0.64 (6.53)	0.14 (14.07)
0.5	0.1	0.9	0.83 (3.71)	0.88 (4.96)	0.04 (17.00)	0.75 (4.90)	0.01 (18.25)
0.5	0.5	0.1	0.96 (1.59)	0.94 (2.03)	0.99 (1.19)	0.99 (1.19)	0.99 (1.19)
0.5	0.5	0.5	0.93 (3.40)	0.93 (3.42)	0.83 (5.74)	0.91 (3.58)	0.79 (6.71)
0.5	0.5	0.9	0.90 (3.87)	0.90 (3.87)	0.09 (21.47)	0.86 (4.03)	0.03 (22.84)
0.5	0.9	0.1	0.60 (7.12)	0.58 (7.49)	0.99 (1.19)	0.68 (6.01)	0.99 (1.19)
0.5	0.9	0.5	0.54 (5.68)	0.54 (6.10)	0.91 (4.01)	0.58 (7.52)	0.90 (5.03)
0.5	0.9	0.9	0.80 (5.00)	0.81 (4.83)	0.20 (34.52)	0.79 (5.58)	0.05 (44.05)
0.9	0.1	0.1	0.74 (10.81)	0.76 (11.66)	0.94 (2.29)	0.86 (6.86)	0.94 (2.43)
0.9	0.1	0.5	0.79 (7.79)	0.86 (4.21)	0.19 (13.26)	0.66 (8.75)	0.16 (13.67)
0.9	0.1	0.9	0.76 (9.28)	0.81 (5.56)	0.04 (15.81)	0.66 (9.20)	0.01 (16.16)
0.9	0.5	0.1	0.80 (9.59)	0.75 (10.74)	0.98 (1.59)	0.88 (6.80)	0.98 (1.59)
0.9	0.5	0.5	0.78 (10.35)	0.78 (10.39)	0.84 (3.91)	0.78 (10.73)	0.80 (4.47)
0.9	0.5	0.9	0.85 (5.50)	0.84 (5.69)	0.05 (16.59)	0.83 (5.48)	0.01 (17.16)
0.9	0.9	0.1	0.61 (8.58)	0.56 (10.99)	0.98 (1.62)	0.68 (7.24)	0.98 (1.62)
0.9	0.9	0.5	0.55 (13.33)	0.55 (10.51)	0.98 (1.41)	0.61 (11.54)	1.0 (1.00)
0.9	0.9	0.9	0.80 (6.80)	0.80 (6.80)	0.61 (11.80)	0.80 (6.77)	0.49 (16.15)

Two statistics are presented: the fraction of times out of 80 the nearest marker exhibited maximum disequilibrium and, in parentheses, the mean-square error times 10^6. Ideally, (10^6) MSE would equal 1.0 because the recombinational distance between the disease locus and the nearest marker was .001.

in complete disequilibrium with other nearby marker loci, our analyses suggest that δ, the robust version of the population attributable risk, is the best measure of disequilibrium for simple fine mapping. From deterministic calculations, it is clear that δ is directly related to θ, the recombination fraction. It is also most closely related to θ for simulations of short-term evolution. Under a more limited set of circumstances, Lewontin's D' yields results comparable to δ. The fact that the two measures behave so similarly, at least under random sampling, is hardly surprising because we have shown that the two are equivalent when the disease is uncommon and marker frequencies are relatively more common in the population. An important caveat is that the measures are not equal when the study, by design, employs case-control sampling.

Alternatively, Δ and d are useful only for simple disequilibrium mapping when marker allele frequencies vary very little from locus to locus, a circumstance unlikely to exist in general. Q is a better measure to use, at least relative to Δ and d. Nevertheless, like Δ and d, marker allele frequency variation across loci has a substantial impact on the pattern of disequilibrium values, especially when some marker allele frequencies are small.

Jorde *et al.* (1994) used Δ to examine the relationship between linkage disequilibrium and physical distance in the adenomatous polyposis coli region. In that study, D' was also examined but the results were not reported; nevertheless, they did report that the values for D' exhibited a pattern similar to those using Δ. It is important to note, however, that the similarity is most likely due to the striking similarity of allele frequencies at the different marker loci rather than to the inherent features of the measures.

The short-term evolutionary simulations that we performed make it clear that forces such as drift, as well as random recombination, influence the relationship between linkage disequilibrium and θ. As shown by Hill and Weir (1994) for steady-state populations, it is apparent that drift can obscure the predicted relationship between recombination fraction and disequilibrium that is critical for simple disequilibrium mapping. In this regard, the MSE statistics from the evolutionary simulations provide a ballpark estimate of the magnitude of error that could be incurred by using simple disequilibrium mapping. However, as we have shown, the performance of simple disequilibrium mapping is affected by variation in marker allele frequencies and by the configuration of markers surrounding the disease locus. Undoubtedly, the MSE is also affected by the amount of time since the initial disease mutation, mutations at marker loci, and so on. In fact, recurrent mutation at marker loci can have a tremendous impact on simple disequilibrium mapping because it mimics recombination.

In this paper, we have focused on disease loci having a single disease allele rather than disease loci with multiple alleles. In addition, we have focused on diseases that are relatively uncommon in the population. From the theory, we see no obvious impact of a common disease on the performance of simple disequilibrium mapping using δ because δ should still be directly related to θ as long as there is only a single disease allele. Of course, the interpretation of δ as a robust approximation to the population attributable risk is questionable because the approximation is poor under these circumstances.

The presence of multiple disease alleles diminishes the strength of the relationship between the disease and the marker alleles. Suppose there are two or more disease alleles, with only proportion α of cases attributable to the primary disease allele. Then

$$\delta' = \frac{\alpha\pi_{1|+1} + (1-\alpha)\pi_{1|+2} - \pi_{1|+2}}{\pi_{2|+2}} = \alpha\left(\frac{\pi_{1|+1} - \pi_{1|+2}}{\pi_{2|+2}}\right) = \alpha\delta,$$

so the association is reduced whenever $\alpha < 1$ (see also Lehesjoki *et al.*, 1993). Since α is constant across loci, δ still gives a consistent pattern of disequilibrium across loci, in theory reaching a maximum at the disease locus. However, the smaller the value of α, the greater the impact of nonsystematic variation (such as drift), reducing localizing power. Nevertheless, our results suggest that an investigator attempting simple fine mapping will generally be most successful when using δ (or possibly D') to describe linkage disequilibrium.

APPENDIX

The evolutionary simulations were performed as follows. First a population of 2000 chromosomes was created that had 20 chromosomes ($p = 0.01$) bearing the disease allele. The alleles at each marker locus on the 20 disease chromosomes were all identical; thus, there was initially complete disequilibrium between disease and marker loci. For the remaining 1980 normal chromosomes, marker alleles were assumed to be independent. The standard conditional independence model was therefore used to create the appropriate number of haplotypes of each possible kind, based on expected frequencies of two-locus haplotypes. (The expected frequencies of normal chromosomes are directly calculable from the marker locus allele frequencies and the observation that $p_{11} = 0.01$ and $p_{21} = 0$ under complete disequilibrium.)

At each generation, the population grew at exponential rate $r = 1.0041607$. To accomplish this growth, a pair of haplotypes was chosen at random from the population at generation t using a standard random number generator. The pair recombined randomly over any of the three intralocus intervals with probability equal to the recombination fractions between loci: 0.001, 0.003, 0.003. Consequently there was no interference between intervals. Two haplotypes were produced by this mechanism, usually the same as before, and then one of them was chosen at random to be a member of the $t + 1$ generation. This procedure was executed n_t^r times to produce the $t + 1$ generation. This method is not completely true to population evolution, in which haplotypes occur in pairs for each person, and only those pairs can recombine. Our procedure, however, is essentially identical to the population process because recombination events are rare, while the simplification allowed us to lower drastically the RAM needed to complete the simulations.

If p dropped below 0.005 at any generation, the simulation was reinitialized at $t = 0$ and run again. This rule kept the frequency of the disease allele from drifting to zero, especially in early generations; of course, a population without disease alleles would not be useful for disequilibrium mapping. On the other hand, marker allele frequencies were allowed to go to fixation. In this (rare) instance, the disequilibrium between the marker and the disease locus was set to zero.

A few special circumstances should be noted: δ should always be positive, so allele labels were reversed whenever necessary, although this rarely occurred; the same action was performed on d; and Q was set to -1 or 1, whichever was appropriate, when one or more cell frequencies were zero. Because the sign of most of the measures is arbitrary, we compared absolute values to determine the maximum disequilibrium value.

At $t = 100$, the entire population was assayed for the patterns of linkage disequilibrium. Then a subsample of 200 disease haplotypes and 200 normal haplotypes were chosen at random, and this sample was analyzed for the patterns of linkage disequilibrium. Several statistics were then recorded. These statistics are discussed in the text.

ACKNOWLEDGMENT

This work was supported in part by NIH Grant HG00348 to N.R.

REFERENCES

Bengtsson, B. O., and Thomson, G. (1981). Measuring the strength of associations between HLA antigens and diseases. *Tissue Antigens* **18:** 356–363.

Boehnke, M. (1994). Limits of resolution of genetic linkage studies: Implications for the positional cloning of human genetic diseases. *Am. J. Hum. Genet.* **55:** 379–390.

Bowcock, A. M., Tomfohrde, J., Weisenbach, J., Bonne-Tamir, B., St. George-Hyslop, P., Giagheddu, M., Cavalli-Sforza, L. L., and Farrer, L. A. (1994). Refining the position of Wilson's disease by linkage disequilibrium with polymorphic microsatellites. *Am. J. Hum. Genet.* **54:** 79–87.

Breslow, N., and Day, N. E. (1980). "Statistical Methods in Medical Research," Vol. I, "The Analysis of Case-Control Studies." IARC, Lyon.

Clegg, M. T., Kidwell, J. F., Kidwell, M. G., and Daniel, N. J. (1976). Dynamics of correlated genetic systems. I. Selection in the region of the Glued locus of *Drosophila melanogaster. Genetics* **83:** 793–810.

Daiger, S. P., Reed, L., Huang, S-S., Zeng, Y-T., Wang, T., Lo, W. H. Y., Okano, Y., Hase, Y., Fukuda, Y., Oura, T., *et al.* (1989). Polymorphic DNA haplotypes at the phenylalanine hydroxylase (PAH) locus in Asian families with phenylketonuria (PKU). *Am. J. Hum. Genet.* **45:** 319–324.

Edwards, A. W. F. (1963). The measure of association in a 2 × 2 table. *J. R. Stat. Soc.* **A126:** 109–114.

Fujita, R., Hanauer, A., Sirugo, G., Heilig, R., and Mandel, J. L. (1990). Additional polymorphisms at marker loci D9S5 and D9S15 generates extended haplotypes in linkage disequilibrium in Friedreich ataxia. *Proc. Natl. Acad. Sci. USA* **87:** 1796–1800.

Graeber, M. B., Kupke, K. G., and Muller, U. (1992). Delineation of the dystonia-parkinsonism syndrome locus in Xq13. *Proc. Natl. Acad. Sci. USA* **89:** 8245–8248.

Haberman, S. J. (1973). The analysis of residuals in cross-classification tables. *Biometrics* **29:** 205–220.

Hanauer, A., Chery, M., Fujita, R., Driesel, A. J., Gilgenkrantz, S., and Mandel, J. L. (1990). The Friedreich ataxia gene is assigned to chromosome 9q13–q21 by mapping of tightly linked markers and shows linkage disequilibrium with D9S15. *Am. J. Hum. Genet.* **46:** 133–137.

Harley, H. G., Brook, J. D., Floyd, J., Rundle, S. A., Crow, S., Walsh, K. V., Thibault, M-C., Harper, P. S., and Shaw, D. J. (1991). Detection of linkage disequilibrium between the myotonic dystrophy locus and a new polymorphic DNA marker. *Am. J. Hum. Genet.* **49:** 68–75.

Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., and Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. *Nature Genet.* **2:** 204–211.

Hästbacka, J., de la Chapelle, A., Mahanti, M. M., Clines, G., Reeve-Daly, M. P., Daly, M., Hamilton, B. A., Kusumi, K., Trivedi, B., Weaver, A., Coloma, A., Lovett, M., Buckler, A., Kaitila, I., and Lander, E. S. (1994). The diastrophic dysplasia gene encodes a novel sulfate transporter: Positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78:** 1073–1087.

Hedrick, P. W. (1987). Gametic disequilibrium measures: Proceed with caution. *Genetics* **117:** 331–341.

Hellsten, E., Vesa, J., Speer, M. C., Makela, T. P., Jarvela, I., Alitalo, K., Ott, J., Peltonen, L. (1993). Refined assignment of the infantile neuronal ceroid lipofuscinosis (INCL, CLN1) locus at 1p32: incorporation of linkage disequilibrium in multipoint analysis. *Genomics* **16:** 720–725.

Hill, W. G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38:** 226–231.

Hill, W. G., and Weir, B. S. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.* **54:** 705–714.

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72:** 971–983.

Jorde, L. B., Watkins, W. S., Viskochil, D., O'Connell, P., and Ward, K. (1993). Linkage disequilibrium in the neurofibromatosis 1 (NF1) region: Implications for gene mapping. *Am. J. Hum. Genet.* **53:** 1038–1050.

Jorde, L. B., Watkins, W. S., Carlson, M., Groden, J., Albertsen, H., Thliveris, A., and Leppert, M. (1994). Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. *Am. J. Hum. Genet.* **54:** 884–898.

Kaplan, N., and Weir, B. S. (1992). Expected behavior of conditional linkage disequilibrium. *Am. J. Hum. Genet.* **51:** 333–343.

Kaplan, N., Hill, W. G., and Weir, B. S. (1995). Likelihood methods for locating disease genes in nonequilibrium populations. *Am. J. Hum. Genet.* **56:** 18–32.

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M., and Tsui, L-C. (1989). Identification of the cystic fibrosis gene: Genetic analysis. *Science* **245:** 1073–1080.

Lehesjoki, A-E., Koskiniemi, M., Norio, R., Tirrito, S., Sistonen, P., Lander, E., and de la Chapelle, A. (1993). Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: Linkage disequilibrium allows high resolution mapping. *Hum. Mol. Genet.* **2:** 1229–1234.

Levin, M. L. (1953). The occurrence of lung cancer in man. *Acta Unio. Int. Contra Cancrum* **19:** 531–541.

Levin, M. L., and Bertell, R. (1978). Re: "Simple estimation of population attributable risk from case-control studies." *Am. J. Epidemiol.* **108:** 78–79.

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

Lewontin, R. C. (1988). On measures of gametic disequilibrium. *Genetics* **120:** 849–852.

Litt, M., and Jorde, L. B. (1986). Linkage disequilibrium between pairs of loci within a highly polymorphic region of chromosome 2Q. *Am. J. Hum. Genet.* **39:** 166–178.

Nei, M. (1987). "Molecular Evolutionary Genetics," Columbia Univ. Press, New York.

Nei, M., and Li, W-H. (1980). Non-random association between electromorphs and inversion chromosomes in finite populations. *Genet. Res.* **35:** 65–83.

Olson, J. M., and Wijsman, E. M. (1994). Design and sample-size considerations in the detection of linkage disequilibrium with a disease locus. *Am. J. Hum. Genet.* **55:** 574–580.

Ott, J. (1991). "Analysis of Human Genetic Linkage," The Johns Hopkins Univ. Press, Baltimore.

Ozelius, L. J., Kramer, P. L., deLeon, D., Risch, N., Bressman, S. B., Schuback, D. E., Brin, M. F., Kwaitkowski, D. J., Burke, R. E., Gusella, J. F., Fahn, S., and Breakefield, X. O. (1992a). Strong allelic association between the torsion dystonia gene (DYT1) and loci in chromosome 9q34 in Ashkenazi Jews. *Am. J. Hum. Genet.* **50:** 619–628.

Ozelius, L. J., Teitz, S. S., Buckler, A., Hervitt, J., Gasser, T., deLeon, D., Kramer, P. L., Risch, N., Bressman, S. B., Housman, D., Fahn, S., Gusella, J. F., and Breakefield, X. O. (1992b). Fine mapping of the torsion dystonia gene (DYT1) on 9q34 and evaluation of a candidate gene. *Am. J. Hum. Genet.* **51**(Suppl.): A224.

Pandolfo, M., Sirugo, G., Antonelli, A., Weirnauer, L., Ferretti, L., Leone, M., Dones, I., Cerino, A., Fujita, R., Hanauer, A., *et al.* (1990). Friedrich's ataxia in Italian families: Genetic homogeneity and linkage disequilibrium with the marker loci D9S5 and D9S15. *Am. J. Hum. Genet.* **47:** 228–235.

Petrukhin, K., Fischer, S. G., Pirastu, M., Tanzi, R. E., Chernov, I., Devoto, M., Brzustowicz, L. M., Cayanis, E., Vitale, E., Russo, J. J., *et al.* (1993). Mapping, cloning and genetic characterization of the region containing the Wilson disease gene. *Nature Genet.* **5:** 338–343.

Richter, A., Morgan, J. K., Poirier, J., Mercier, J., Chamberlain, S., Mandel, J. L., and Melancon, S. B. (1990). Friedreich's ataxia: Linkage disequilibrium in the Quebec French Canadian population. *Am. J. Hum. Genet.* **47**(Suppl.): A144.

Riordan, J. R., Rommens, H. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J-L., Drumm, M. L., Iannuzzi, M. C., Collins, F. S., and Tsui, L-C. (1989).

Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science* **245:** 1066–1073.

Risch, N., deLeon, D., Ozelius, L. J., Kramer, P., Bressman, S., Kwiatkowski, D., Brin, M. F., *et al.* (1991). The genetics of torsion dystonia in Ashkenazi Jews. *In* "Molecular Genetics and Neuropsychiatric Disorders," p. A6, Scientific Program and Abstracts, Israel Ministry of Science and Technology.

Risch, N., deLeon, D., Ozelius, L. J., Kramer, P., Almasy, L., Singer, B., Fahn, S., Breakefield, X., and Bressman, S. (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent decent from a small founder population. *Nature Genet.* **9:** 152–159.

Rommens, H. M., Iannuzzi, M. C., Kerem, B., Drumm, M. L., Melmer, G., Dean, N., Rozmahel, R., Cole, J. L., Kennedy, D., Hidaka, N., Zsiga, M., Buchwald, M., Riordan, J. R., Tsui, L-C., and Collins, F. S. (1989). Identification of the cystic fibrosis gene: Chromosome walking and jumping. *Science* **245:** 1059–1065.

Shiang, R., Thompson, L. M., Zhu, Y-Z., Church, D. M., Fiedler, T. J., Bocian, M., Winokur, S. T., and Wasmuth, J. J. (1994). Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, Achondroplasia. *Cell* **78:** 335–342.

Sirugo, B., Keats, B., Fujita, R., Duclos, F., Purohit, K., Koenig, M., and Mandel, J. L. (1992). Friedreich ataxia in Louisiana Acadians: Demonstration of a founder effect by analysis of microsatellite-generated extended haplotypes. *Am. J. Hum. Genet.* **50:** 559–566.

Snarey, A., Thomas, S., Schneider, M. C., Pound, S. E., Barton, N., Wright, A. F., Somlo, S., Germino, G. G., Harris, P. C., Reeders, S. T., and Frischauf, A. M. (1994). Linkage disequilibrium in the region of the autosomal dominant polycystic kidney disease gene (PKD1). *Am. J. Hum. Genet.* **55:** 365–371.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic loci. *Am. J. Hum. Genet.* **56:** 777–787.

Thompson, E. A., Deeb, S., Walker, D., and Motulsky, A. G. (1988). The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. *Am. J. Hum. Genet.* **42:** 113–124.

Thomson, G. (1981). A review of theoretical aspects of HLA and disease associations. *Theor. Pop. Biol.* **20:** 168–208.

Tsilfidis, C., McKenzie, A. E., Shutler, G., Leblond, S., Bailly, J., Johnson, K., Williamson, R., Siegel-Bartelt, J., and Korneluk, R. G. (1991). D17S51 is closely linked with and maps distal to the myotonic dystrophy locus on 19q. *Am. J. Hum. Genet.* **49:** 961–965.

Walter, M. A., and Cox, D. W. (1991). Nonuniform linkage disequilibrium within a 1,500-kb region of the human immunoglobulin heavy-chain complex. *Am. J. Hum. Genet.* **49:** 917–931.

Walter, S. D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika* **62:** 371–374.

Weir, B. S. (1989). Locating the cystic fibrosis gene on the basis of linkage disequilibrium with markers? *In* "Multipoint Mapping and Linkage Based on Affected Pedigree Members: Genetic Analysis Workshop 6" (R. C. Elston, M. A. Spence, S. E. Hodge, J. W. MacCluer, Eds.), pp. 81–86, A. R. Liss, New York.

Weir, B. S. (1990). "Genetic Data Analysis," Sinauer, Sunderland, MA.

Wilhelmsen, K. C., Weeks, D. E., Neystat, M., and Nygaard, T. G. (1992). Linkage disequilibrium mapping of lubag (X-linked dystonia-parkinsonism) using simple sequence repeats. *Am. J. Hum. Genet.* **51**(Suppl.): A205.

Yule, G. U. (1900). On the association of attributes in statistics. *Philos. Trans. R. Soc. London A* **194:** 257–319.