*Am. J. Hum. Genet. 51:333–343, 1992*

# Expected Behavior of Conditional Linkage Disequilibrium

Norman Kaplan* and B. S. Weir†

*Division of Biometry and Risk Assessment, National Institute of Environmental Health Sciences, Research Triangle Park, NC; and †Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh

## Summary

**The ubiquitousness of RFLPs in the human genome has greatly helped the mapping of human disease genes, and it has been suggested that population measures of association between disease and marker loci could help with this mapping. For rare diseases, random samples are taken from within disease genotypes in order to obtain reasonable sample sizes, but this sampling strategy requires a modification of the usual measures of association. We present theoretical predictions for the mean and variance of such a modified measure, under the assumption that the disease gene is maintained at a constant low frequency in the population. The coefficient of variation of this modified measure is large enough that caution is needed in using the measure to locate disease genes, and, furthermore, the coefficient of variation cannot be made arbitrarily small by increasing sample size. The modified association measure is calculated for recently published data on cystic fibrosis.**

## Introduction

The discovery of widespread existence RFLPs has greatly accelerated the mapping of the human genome. As envisioned by Botstein et al. (1980), the mapping procedure depends on the determination, throughout the genome, of a large number of well-spaced RFLPs. These marker loci can be used as reference points on the genetic map, and a gene of interest can be localized by identifying the markers to which it is closely linked. This genetic approach is appropriate for diseases in which the gene product is not known and in which direct probing of the genome is not possible. The approach is currently being implemented with some success, and the chromosomal locations of a number of disease genes have already been reported (e.g., see Gusella et al. 1983; Eiberg et al. 1985; Kerem et al. 1989).

The chromosomal density of RFLPs is an important determinant of how well a disease gene can be localized. Even if a known RFLP is within a centiMorgan (cM) of the gene, it is still necessary to search over a million base pairs to determine exact location and begin a molecular characterization of the gene. In practice, it is not known a priori which markers are close to the disease locus, and it is often suggested (e.g., see Pritchard et al. 1991) that population measures of association (or linkage disequilibrium) between disease and marker can be used to identify the most likely chromosome region where the disease gene may lie. Any such clue as to the position of a locus, particularly if it was more precise than can be achieved from linkage studies with family data, would be of considerable help in molecular studies.

Identifying genomic regions on the basis of statistical measures of association does not depend on any population genetic theory or model, other than that which supposes the strength of association between genes to increase as physical and genetic distances between them decrease. An obvious question to explore is what could be gained by invoking a population genetic model. We recently asked this question with reference to cystic fibrosis (CF) (Weir 1989), but our concerns about the large sampling variance of measures of association (Hill and Weir 1988) may have been influenced by our analysis not taking account of

Received October 24, 1991; final revision received March 9, 1992.

Address for correspondence and reprints: Norman Kaplan, Division of Biometry and Risk Assessment, NIEHS, P.O. Box 12233, Research Triangle Park, NC 27709.

0002-9297/92/5102-0013$02.00

the conditional sampling structure of the data in those studies.

Previous work by ourselves and others (e.g., see Hill and Weir 1988) was based on moments of gene frequencies at two loci, with attention paid to the steady-state expectation of measures of association under models of drift, recombination, mutation, and selective neutrality. Expectations were expressed in terms of a parameter, $\Gamma = 4Nr$, where $N$ is the diploid population size and where $r$ is the recombination rate between the two genes. It has been proposed that this parameter could be estimated by equating observed and expected values of the association measure (e.g., see Estivill et al. 1987). Furthermore, if it is assumed that physical and genetic distances are linearly related, which is not unreasonable for recombination fractions less than 1 cM, then the physical distance between marker and disease loci could be inferred from the known physical distances between markers.

We have noted (Weir and Hill 1986; Hill and Weir 1988; Weir 1989) that a major problem with this approach arises with the large variances of measures of association. The sampling-variance component of the total variance can be minimized with sufficiently large samples, but the genetic variance arising from the stochastic nature of evolutionary forces is beyond the control of the investigator.

The expected values of the measures of association discussed by Hill and Weir (1988) were derived under the assumption that individuals were sampled at random from the population. For rare diseases, of course, samples are taken independently from within each disease genotype category in order to obtain sufficiently large samples. The CF data discussed by Weir (1989) were, in fact, for disease-locus heterozygotes only (Weir 1990), and measures of association for such a sampling scheme have been discussed (Chakravarti et al. 1984; Chakraborty 1986).

A second limitation of the traditional approach is that it is based on expectations over all replicate populations, whereas the data analyzed in practice require at least minimal levels of polymorphisms at the marker loci. Expectations conditional on polymorphism were investigated by Hudson (1985), for random samples of chromosomes at selectively neutral loci. On the basis of simulation results, Hudson showed that these conditional expectations were substantially larger than unconditional expectations. Although he did not present variance calculations, his work suggests that sampling variance may not be as important an issue for conditional as it is for unconditional measures.

The conventional measure of linkage disequilibrium is not appropriate for samples collected from each disease genotype. A more suitable measure was proposed by Nei and Li (1980) and Chakravarti et al. (1984). In this paper we present the predicted mean and variance of this measure, under the assumption that the disease gene frequency in the population remains more or less constant over time. Our work depends on coalescent theory for models with balancing selection (Kaplan et al. 1988). We also present a reanalysis of the CF data of Kerem et al. (1989).

## Theory

Suppose the disease is caused by a single gene H. Evolutionary forces are assumed to act to preserve an approximately constant frequency of the disease in the population over time. Possible mechanisms are high mutation rate, increased fertility, reproductive compensation, and heterozygote advantage. A discussion of these mechanisms for cystic fibrosis was given by Tsui and Buchwald (1991, p. 233). If there is a single disease allele $H$, then its frequency $q$ will also be approximately constant. Diseases with a low mutation rate are likely to have arisen from few, or possibly one, mutation. Huntington disease appears to fall into this category (Gusella 1991). Also, the strong linkage disequilibrium between the CF locus and closely linked markers suggests that most, if not all, CF genes are descended from a single ancestral mutation (Tsui and Buchwald 1991). The discovery that approximately 70% of the CF genes carry the same 3-bp deletion (Kerem et al. 1989) supports this hypothesis. The assumption of a single disease allele may not be correct, but it is the parsimonious one if haplotypes are characterized only as being either diseased or normal. This issue will be taken up in the Discussion. The normal allele $h$ has frequency $p$ ($p + q = 1$). For rare diseases, $q$ will be small, and it is about .02 for CF (Kerem et al. 1989).

Suppose marker locus M, with segregating alleles $M_1$ and $M_2$, is linked to H. If the four haplotypes $HM_1$, $HM_2$, $hM_1$, and $hM_2$ have population frequencies $X_1$, $X_2$, $X_3$, and $X_4$, respectively, then the frequencies of allele $M_1$ among disease and normal haplotypes are $x = f(M_1|H) = X_1/q$ and $y = f(M_1|h) = X_3/p$, the coefficient of linkage disequilibrium $D$ in the population is $D = f(HM_1) - f(H)f(M_1) = pq(x - y)$, and the squared correlation of gene frequencies, $r^2$, is $r^2 = D^2/pqm_1m_2$, where $m_1$ and $m_2$ are the population frequencies of marker alleles $M_1$ and $M_2$, respectively.

Since $pq$ is assumed to be constant, a measure of association for data conditional on disease locus status is $d = x - y$. In this paper we are concerned with the magnitude of $d$ and not with its sign, so it is more convenient to work with $d^2$. Nei and Li (1980) studied the transient behavior of expected values of $d$ and $d^2$ for models without recurrent mutation. Estimation procedures from genotypic data have been discussed by Chakravarti et al. (1984) and Maiste and Weir (1992). Our goal is to calculate the equilibrium mean and variance of the distribution of $d^2$ under the assumption that variation at the marker locus follows a selectively neutral infinite-site model with recurrent mutation (Kimura 1969, 1971).

For a selectively neutral model with recurrent mutation, Hill and Weir (1988) expressed the moments of $D^2$ as linear combinations of identity coefficients, and their approach can be extended to apply to $d^2$. We define the identity coefficient $\phi$ as $\phi(i,j) = \Pr(i$ randomly sampled haplotypes from the subpopulation carrying the $h$ allele and $j$ randomly sampled haplotypes from the subpopulation carrying the $H$ allele have the same allele at the marker locus).

From the definitions of the conditional frequencies $x$ and $y$, it follows that $\phi(i,j) = \varepsilon[x^i y^j + (1 - x)^i (1 - y)]^j$, $i,j \geqslant 0$, where expectation, denoted by $\varepsilon$, is over all replicate populations subject to the same forces. Independence of haplotypes, or random mating, is assumed.

Expanding the definition of $d^2$ allows its expected value to be written in terms of identity coefficients: $2d^2 = 2(x^2 - 2xy + y^2) = [x^2 + (1 - x)^2] - 2[xy + (1 - x)(1 - y)] + [y^2 + (1 - y)^2]$, so that

$$2\varepsilon(d^2) = \phi(2,0) - 2\phi(1,1) + \phi(0,2) . \qquad (1)$$

Similarly,

$$2\varepsilon(d^4) = \phi(4,0) - 4\phi(3,1) + 6\phi(2,2) - 4\phi(1,3) + \phi(0,4) . \qquad (2)$$

To calculate the mean and variance of $d^2$, it is therefore sufficient to calculate the identity coefficients $\phi(i,j)$ for appropriate values of $i$ and $j$.

In practice, only estimates of $d^2$ are available. For a sample of haplotypes, $n_H$ with the disease gene and $n_{HM_1}$ carrying the $M_1$ marker, and $n_h$ with the normal gene and $n_{hM_1}$ carrying the $M_1$ marker, the maximum-likelihood estimate of $d^2$ is $\hat{d}^2 = [(n_{HM_1}/n_H) - (n_{hM_1}/n_h)]^2$. A slight extension of the use of identity coefficients can be used to compute moments of such estimates. Following the same reasoning as used by Hudson (1985), we find that the first and second moments of $\hat{d}^2$ satisfy equations (1) and (2), providing that the $\{\phi(i,j)\}$ are replaced by $\{\phi^*(i,j)\}$, shown in the Appendix.

The formulas for the moments of $\hat{d}^2$ are not strictly appropriate for interpreting data, since the possibility of the sample of haplotypes being monomorphic at the marker locus was not excluded from the derivations. Only polymorphic markers are used, and this needs to be taken into account. Minimally, there is a need for information about the distribution of $\hat{d}^2$ conditional on there being at least two different marker alleles in the sample. Although it would be more informative to condition on the marker sample frequencies, this distribution is very difficult to obtain. Hudson (1985) used a simulation technique to find this conditional distribution for two linked neutral loci.

Let $k_M$ be the number of different alleles in the sample at the marker locus. Since $\hat{d}^2 = 0$ for $k_M = 1$,

$$\varepsilon(\hat{d}^2|k_M > 1) = \frac{\varepsilon(\hat{d}_2)}{\Pr(k_M > 1)} = \frac{\varepsilon(\hat{d}^2)}{1 - \phi(n_H, n_h)} . \qquad (3)$$

A similar result holds for the fourth moment. The conditional moments are therefore easily obtained from the unconditional moments. To simplify notation, the moments of $\hat{d}^2$ conditional on polymorphism will be subscripted by $P$.

Further assumptions about the evolutionary forces acting on the population and about the nature of the variation at the marker locus are needed before the identity coefficients can be calculated. We suppose the population is finite, of size $N$, mating at random, and that the marker locus variation is described by a selectively neutral infinite-site model (Kimura 1969, 1971; Watterson 1975). More specifically, mutations occur at rate $\mu$ per locus per generation, and each mutant is unique. We also assume that there is no recombination within the locus but that there is recombination between marker and disease loci.

Suppose two haplotypes are sampled at random, and ignore for the moment which allele is present at the disease locus. Under a selectively neutral infinite-site model, the two marker genes are the same allele if and only if there has been no mutation among their ancestral genes since their most recent common ancestor (MRCA). If $T_0$ is the ancestral generation containing the MRCA, and if $G_2$ is the probability of the two marker genes having the same allele, then

$$\begin{aligned} G_2 &= \varepsilon[(1-\mu)^{T_0}(1-\mu)^{T_0}] \\ &\approx \varepsilon(e^{-2\mu T_0}) \qquad (4) \\ &= \varepsilon(e^{-\frac{\theta T}{2}}), \end{aligned}$$

where $\theta = 4N\mu$ and $T = 2T_0/2N$ (i.e., time is now measured in units of $2N$ generations).

It is not difficult to generalize this argument to larger samples of haplotypes. If $G_k$ is the probability that the marker genes on $k$ sampled haplotypes are identical by descent, then

$$G_k = \varepsilon(e^{-\frac{\theta T}{2}}), \qquad (5)$$

where $T$ is now the size of the ancestral tree describing the genealogical history of the size-$k$ sample at the M locus, measured in units of $2N$ generations.

The ancestral tree keeps track of when each common ancestor of the sampled marker genes occurs, as well as which two sampled genes have the common ancestor. In figure 1 we show the ancestral tree for a sample of size 4. A more detailed discussion is given by Kaplan et al. (1988).

A direct consequence of the selective neutrality of the marker locus is that equation (5) continues to hold even conditioning on the disease alleles with which the marker alleles are associated. The conditioning affects only the distribution of $T$. To calculate the $\phi(i,j)$'s, therefore, it is sufficient to determine the statistical properties of $T$ conditional on $i$ of the sampled marker genes being on the same haplotypes as are $H$ alleles, with $j$ being haplotypic to $h$ alleles. Recent work of Hudson and Kaplan (1988) shows how these conditional distributions may be calculated for large populations, and these results are now described.

Hudson and Kaplan proved that, if $N$ is large, then the distribution of $T$ can be obtained by considering the behavior of a finite-state Markov chain. Suppose that $n_H$ haplotypes are sampled carrying allele $H$ and

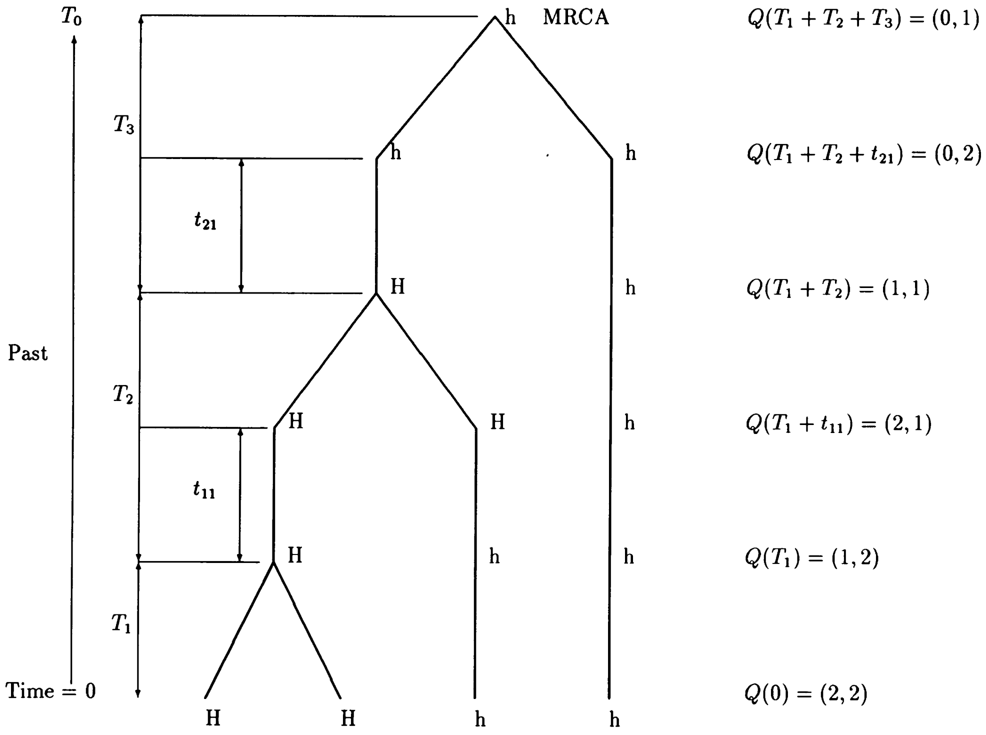

**Figure 1** Ancestral tree for a sample of size four, with two haplotypes carrying $H$ and two carrying $h$. The $Q$ process changes value five times. At ancestral generations $T_1$, $T_1 + T_2$, and $T_1 + T_2 + T_3$, a common ancestor event occurs and $|Q|$ decreases by one. At ancestral generations $T_1 + t_{11}$ and $T_1 + T_2 + t_{21}$, recombination between the marker and disease loci occurs. The MRCA of the sample occurs at ancestral generation $T_0 = T_1 + T_2 + T_3$.

that $n_h$ haplotypes are sampled carrying $h$. For the $t$th ancestral generation, define $Q(t) = (i,j)$ if $i$ of the ancestral genes of the sampled marker genes are linked to an $H$ allele and if $j$ are linked to an $h$ allele (see fig. 1). Further, define $|Q(t)| = i + j$. Since the ancestral genes of the sampled marker genes are necessarily linked to either an $H$ or an $h$, the $Q$ process is well defined, and $Q(0) = (n_H,n_h)$. There are two ways in which the $Q$ process can change its value: $Q(t) = (i,j)$ and $Q(t + 1) = (i - 1,j)$ or $(i,j - 1)$, and $Q(t) = (i,j)$ and $Q(t + 1) = (i - 1,j + 1)$ or $(i + 1,j - 1)$.

The first case corresponds to a common ancestor event. If $i + j = 1$, then all the common ancestors of the sample genes have occurred, and so $T_0 = \min(t\colon |Q(t)| = 1)$ is the ancestral generation in which the most recent common ancestor of the sample occurs. The second case corresponds to a recombination event. For example, if $Q(t) = (i,j)$ and if $Q(t + 1) = (i - 1,j + 1)$, then an $M_1H$ haplotype in the $t$th ancestral generation derives from an $M_1h/M_2H$ parent in the $(i + 1)$th ancestral generation. Recombination events do not alter $|Q(t)|$.

Hudson and Kaplan show that, if time is measured in units of $2N$ generations, then the $Q$ process can be approximated by a continuous-time finite-state Markov process with the following parameters: The time $T_{ij}$ until the $Q$ process changes its state has an exponential distribution with parameter $\lambda_{ij} = i(i - 1)/2q + j(j - 1)/2p + ip\Gamma/2 + jq\Gamma/2$, where $\Gamma = 4Nr$ and $r$ is the recombination rate between the marker and disease loci. Furthermore, when the $Q$ process does change state in the small time interval between $t$ and $t + \delta t$, the probability distribution describing how this change occurs is given by

$$\Pr[Q(t + \delta t) = (i - 1,j)|Q(t) = (i,j)] = \frac{i(i - 1)}{2q\lambda_{ij}}$$

$$\Pr[Q(t + \delta t) = (i,j - 1)|Q(t) = (i,j)] = \frac{j(j - 1)}{2p\lambda_{ij}}$$

$$\Pr[Q(t + \delta t) = (i - 1,j + 1)|Q(t) = (i,j)] = \frac{ip\Gamma}{2\lambda_{ij}}$$

$$\Pr[Q(t + \delta t) = (i + 1,j - 1)|Q(t) = (i,j)] = \frac{jq\Gamma}{2\lambda_{ij}}.$$

The size $T$ of the ancestral tree can be written as

$$T = \sum_{j=2}^{|Q(0)|} jT_j ,$$

where $T_j$ is the length of time during which $|Q(t)| = j$. A more convenient representation of $T$ is

$$T = \int_0^{T_0} |Q(u)|du .$$

The identity coefficient $\phi(i,j)$ can therefore be written as $\phi(i,j) = \varepsilon\left[e^{-\frac{\theta}{2}\int_0^{T_0}|Q(u)|du}|Q(0) = (i,j)\right]$. By means of this representation of $\phi(i,j)$ and the Markov properties of the $Q$ process, it is straightforward to obtain the following general recursion for the identity coefficients:

$$\begin{aligned}\phi(i,j) &= \varepsilon\left[e^{-\frac{\theta}{2}[(i+j)T_{ij}+\int_{T_{ij}}^{T_0}|Q(u)|du]}|Q(0) = (i,j)\right] .\\ &= \frac{\lambda_{ij}}{\frac{(i+j)\theta}{2}+\lambda_{ij}}\left[\frac{i(i-1)}{2q\lambda_{ij}}\phi(i-1,j)\right.\\ &\quad +\frac{j(j-1)}{2p\lambda_{ij}}\phi(i,j-1)+\frac{ip\Gamma}{2\lambda_{ij}}\phi(i-1,j+1)\\ &\quad \left.+\frac{jq\Gamma}{2\lambda_{ij}}\phi(i+1,j-1)\right] . \qquad (6)\end{aligned}$$

The derivation of equation (6) is obtained by conditioning on the time of the first jump of the $Q$ process and on the state to which the jump is made.

For any sample size $n > 1$, equation (6) results in a system of linear equations in which all of the $\phi(i,j)$ with $i + j < n$ must be evaluated in order to obtain the measure with $i + j = n$. This system of equations is tri-diagonal, and so is easy to solve numerically even for large values of $n$ (Press et al. 1986).

The case of $n = 2$ is simple enough to present as an example:

$$\phi(2,0) = \frac{1}{q(\theta + \lambda_{20})} + \frac{p\Gamma\phi(1,1)}{\theta + \lambda_{20}}$$

$$\phi(0,2) = \frac{1}{p(\theta + \lambda_{02})} + \frac{q\Gamma\phi(1,1)}{\theta + \lambda_{02}}$$

$$\phi(1,1) = \frac{\Gamma}{2(\theta + \lambda_{11})}[p\phi(0,2) + q\phi(2,0)] ,$$

where $\lambda_{20} = (1/q) + p\Gamma$, $\lambda_{02} = (1/p) + q\Gamma$, and $\lambda_{11} = \Gamma/2$. As $\Gamma$ becomes large, simple algebra shows that each of $\phi(2,0)$, $\phi(0,2)$, and $\phi(1,1)$ approaches $1/(1 + \theta)$, the value of homozygosity for a selectively neutral locus (Ewens 1979). Alternatively, as $\Gamma$ approaches 0, $\phi(2,0) \rightarrow 1/(1 + q\theta)$, $\phi(0,2) \rightarrow 1/(1 +$

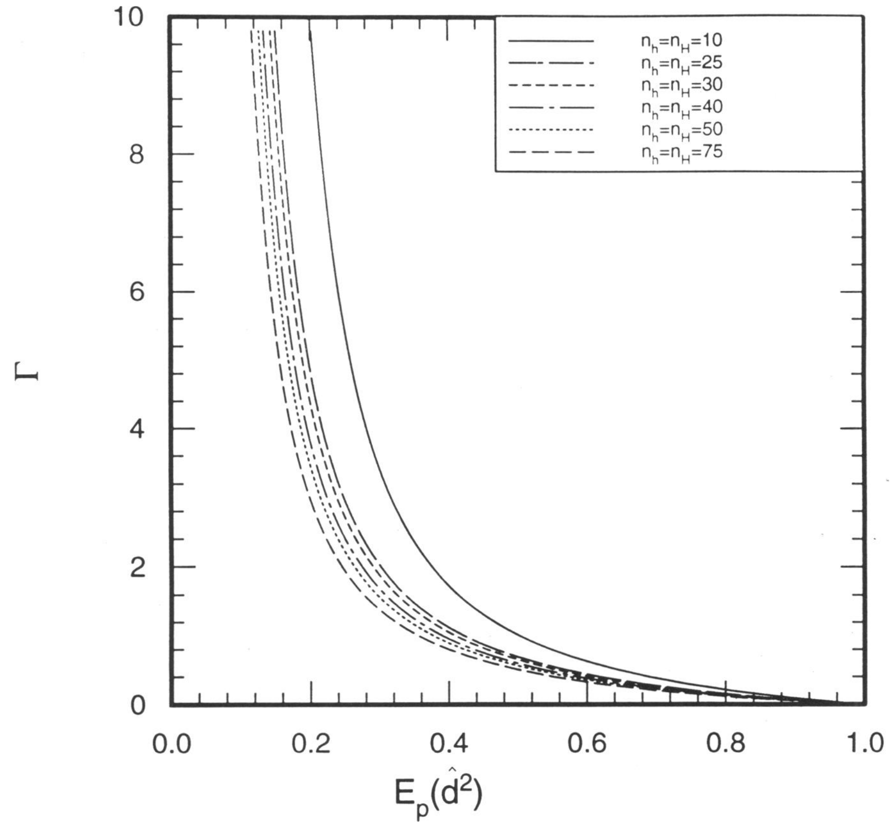

**Figure 2** $\Gamma = 4Nr$ plotted against $\varepsilon_p(\hat{d}^2)$. Each curve is for a different sample size, $q = .02$ and $\theta = .01$

$p\theta$), and $\phi(1,1) \rightarrow 0$, so that, for this case of complete linkage, $\varepsilon(d^2) = \frac{1}{2}\,[1/(1+q\theta) + 1/(1+p\theta)]$ .

### Numerical Calculations

The conditional distribution of $\hat{d}^2$ depends on the parameters $q$, $\theta$, $\Gamma$, $n_H$, and $n_h$. For many data sets, the two sample sizes are approximately the same, so we set $n_H = n_h$. For human diseases, the population frequency of the disease allele is small, and here we take $q = .02$, to allow application to CF (Kerem et al. 1989). Since RFLPs seldom have more than two alleles, we set the mutation parameter $\theta = .01$. However, other calculations have shown us that the results are robust to small changes in $\theta$. It is the recombination parameter $\Gamma$ that is of greatest interest. In figure 2 we plot the conditional mean of $\hat{d}^2$ as a function of $\Gamma$. Figure 2 shows that sample size has a small effect on the magnitude of $\varepsilon_p(\hat{d}^2)$, provided that $n_H$ and $n_h$ are greater than about 25.

The sharp increase in $\Gamma$ as $\varepsilon_p(\hat{d}^2)$ decreases in figure 2 indicates that moment estimates of $\Gamma$ obtained from sample values of $\hat{d}^2$ less than .2 are probably not reliable. For example, if $n_h = n_H = 50$ and the observed value of $\hat{d}^2$ is .2, then the estimate of $\Gamma$ is 3.4. If the value of $\varepsilon_p(\hat{d}^2)$ is .1 and the value $\hat{d}^2 = .2$ occurs by chance, then the correct value of $\Gamma$ is 6.5, which is about twice the estimate. It is therefore important to assess the likelihood of observing .2 when the true value is .1. In figure 3 we plot the conditional coefficient of variation of $\hat{d}^2$ for different sample sizes. This quantity is an increasing function of $\Gamma$, with a maximum value of about 2. If $n_H = n_h = 50$, then the coefficient of variation exceeds 1 for $\Gamma$ values greater than .7, and an observed value of .2 when the true value of the conditional mean is .1 is not unlikely.

Most of the variation is due to genetical sampling (between replicate populations) and therefore cannot be reduced by increasing the sample size. Even though the coefficient of variation of $\hat{d}^2$ is not negligible, it is

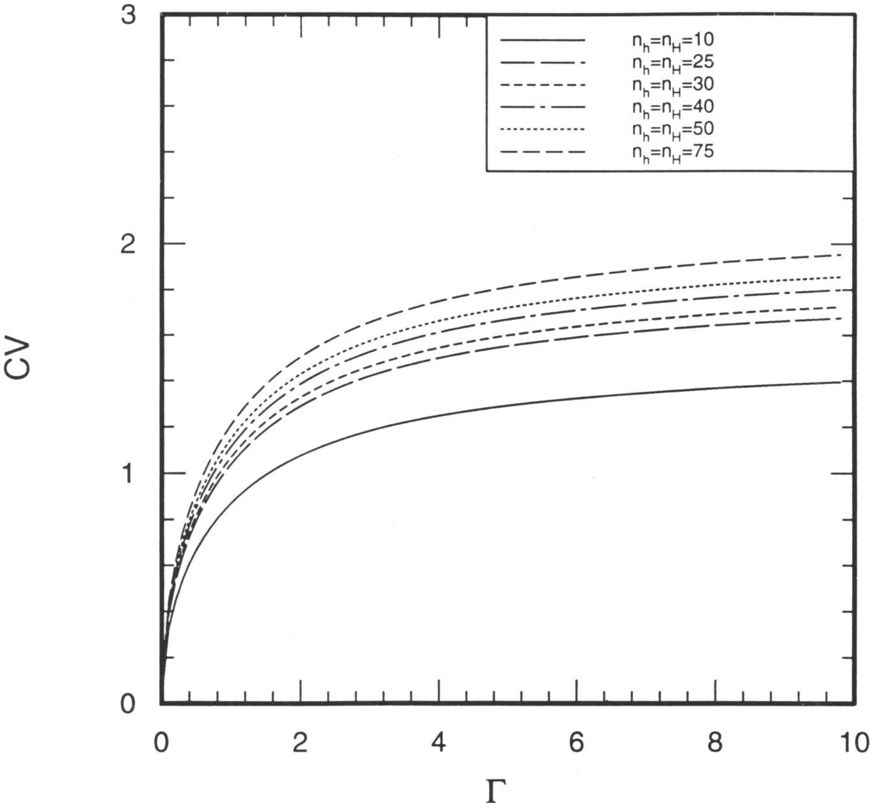

**Figure 3** Coefficient of variation (CV) of the distribution of $\hat{d}^2$, conditional on polymorphism at the marker locus plotted against $\Gamma = 4Nr$. Each curve is for a different sample size, $q = .02$ and $\theta = .01$.

substantially smaller than that of $\hat{r}^2$ (Hill and Weir 1988).

We have also investigated the effects of disease frequency $q$ and mutation rate $\theta$. As long as these quantities are less than .1, which is undoubtedly the case, they have little effect on the conditional moments of $\hat{d}^2$.

## Application

In table 1 we show the quantities $\hat{d}^2$ and $\hat{r}^2$ for the CF data discussed by Weir (1989). Some loci had sample sizes that were judged to be too small for sampling effects to be ignored and so were excluded. Both measures localize the CF locus to the same region of the chromosome, which is not surprising, in view of the algebraic relation between them: $\hat{d}^2 = (n_{M_1}n_{M_2}/n_Hn_h)\hat{r}^2$, where $n_{M_1}$ and $n_{M_2}$ haplotypes in the sample carry the $M_1$ and $M_2$ marker alleles, respectively. For most of the markers the ratio $n_{M_1}n_{M_2}/n_Hn_h$ is close to one. We have obtained distributional results for $d^2$, while previous results for $\hat{r}^2$ apply to random samples and so are not appropriate for these CF data collected only from heterozygotes.

Kerem et al. (1989) published frequency data for 17 markers much closer to the CF locus. The values of $\hat{d}^2$ and the associated estimates of Γ are displayed in table 2. The $\Delta F_{508}$ deletion associated with CF is located between *10–1X.6/HaeIII* and *T6/20/MspI* (Kerem et al. 1989). It is clear from table 2 that the two measures, $\hat{r}^2$ and $\hat{d}^2$, again are about equivalent in their abilities to identify markers close to the disease gene. The results suggest that the CF gene lies in the region flanked by markers *EG1.4/HincII* and *H1.3/NcoI.* Following the suggestion of Estivill et al. (1987), we estimate Γ with the value for which the conditional expectation of $\hat{d}^2$ equals its observed value, under the assumptions $q = .02$ and $\theta = .01$. Markers *10–1.6/AccI* and *10–1X.6/HaeIII* have the smallest Γ estimates (and the highest $\hat{d}^2$ values). It would be expected

## Table 1

**Data from Weir (1989)**

| Probe | Enzyme | $n$ | $\hat{r}^2$ | $\hat{d}^2$ |
|---|---|---|---|---|
| NJ1 | *Msp*I | 146 | .0004 | .0002 |
| NJ3 | *Eco*RI | 122 | .0029 | .0024 |
| 917 | *Hind*III | 128 | .0004 | .0002 |
| 917 | *Hinc*II | 110 | .0030 | .0030 |
| B79a | *Hind*III | 118 | .0000 | .0000 |
| B79a | *Msp*I | 140 | .0039 | .0033 |
| pF167.2 | *Bgl*II | 160 | .0000 | .0000 |
| pB130 | *Taq*I | 150 | .0068 | .0007 |
| pB174 | *Taq*I | 164 | .0022 | .0006 |
| pJB521 | *Sca*I | 166 | .0012 | .0006 |
| pB192 | *Taq*I | 150 | .0035 | .0007 |
| pA37 | *Pst*I | 162 | .0344 | .0344 |
| Z5-1 | *Taq*I | 146 | .0008 | .0008 |
| 7C22 | *Eco*RI | 168 | .0056 | .0035 |
| met5-9 | *Taq*I | 160 | .0069 | .0056 |
| metD | *Ban*I | 164 | .0933 | .0930 |
| metD | *Taq*I | 178 | .0316 | .0153 |
| metH | *Taq*I | 162 | .0194 | .0184 |
| E7 | *Taq*I | 166 | .0630 | .0524 |
| pH-131 | *Hinf*I | 180 | .1919 | .1783 |
| W5-1.4 | *Hind*III | 141 | .1381 | .1296 |
| pJB89 | *Msp*I | 180 | .0036 | .0011 |
| J3.11 | *Msp*I | 166 | .0428 | .0420 |
| J3.11 | *Taq*I | 180 | .0036 | .0011 |
| 311:p3HI | *Pvu*II | 168 | .0074 | .0023 |
| TM183 | *Taq*I | 168 | .0165 | .0023 |
| Z8-2 | *Bgl*II | 176 | .0004 | .0001 |
| SA37 | *Pst*I | 166 | .0123 | .0093 |
| pB-48 | *Sst*I | 134 | .0037 | .0036 |
| SC33 | *Hd*III/*Bgl* | 115 | .0000 | .0000 |
| pB47L | *Bgl*II | 176 | .0173 | .0012 |
| pA51 | *Taq*I | 162 | .0010 | .0006 |
| SI32 | *Msp*I | 118 | .0270 | .0233 |
| SI32 | *Dra*I | 158 | .0175 | .0160 |
| pA21 | *Pst*I | 154 | .0020 | .0007 |
| pG54 | *Pst*I | 144 | .0139 | .0123 |
| TCRB | *Bgl*II | 104 | .0033 | .0033 |

that $\Gamma$ estimates would increase more or less monotonically with distance away from the *10–1X.6* region, but this is not the case. The large predicted values for the coefficient of variation for $\hat{d}^2$, as well as the rapid increase in the value of $\varepsilon_p(\hat{d}^2)$ for small values of $\Gamma$ shown in figure 2, suggest that sampling error is a likely explanation for the nonmonotonicity of the estimated $\Gamma$ values. Another possible explanation is that recombination may not be uniform in the region, and Kerem et al. (1989) suggest that there may be a recombinational hot spot in the vicinity of *J44*.

## Discussion

Attempts to locate human disease genes on the basis of observed linkage disequilibria to markers continue (e.g., see Kupke et al. 1991). These studies should take account of the way in which data were sampled, since, if samples were taken from within disease categories, it is necessary to use conditional measures of linkage disequilibrium. We have given a theoretical treatment of how these measures are expected to behave as functions of the recombination fractions between disease and marker loci, for the case when the disease is controlled by a single locus.

Family-based linkage analysis can localize a disease locus to about 1 cM, which is not sufficient for molecular studies. Although more detailed localizations rest on physical mapping, there may be an advantage to population-based association analyses to aid in the search for genes. The results of Kerem et al. (1989), as emphasized by Pritchard et al. (1991), show that linkage disequilibrium did play a supporting role in the search for the CF gene, while Snell et al. (1989) and Theilmann et al. (1989) showed that linkage disequilibrium was considered in the search for the Huntington disease gene.

Traditional theory for the association measure $\hat{r}^2$ assumes that data have been collected from a random sample from the population as a whole. This is not the case for rare human diseases, where, instead, sampling is done within disease categories. Although this implies that theory for the expected behavior of $\hat{r}^2$ does not apply, the fact the $n\hat{r}^2$ is just the $\chi^2$ test statistic for detecting association between frequencies at two loci means that it remains appropriate to use $\hat{r}^2$ to identify regions of a chromosome that are likely to have a particular disease gene.

For conditional haplotype data, an appropriate measure of association is $\hat{d}^2$. As long as the ratio $n_{M_1}n_{M_2}/n_Hn_h$ is close to 1, $\hat{r}^2$ and $\hat{d}^2$ are approximately equal, and this is the case for the CF data in tables 1 and 2. It is the case because disease-locus heterozygotes are sampled and because markers with a high degree of polymorphism are used.

Since marker loci are necessarily polymorphic, we work with the conditional distribution of $\hat{d}^2$. Figure 2 shows that the conditional mean of $\hat{d}^2$ is large enough so that finite sampling effects are negligible, provided that subsample sizes are 20 or more, which was the case for the CF studies. The coefficient of variation for $\hat{d}^2$ is low for small values of $\Gamma$, and it increases with $\Gamma$, to a maximum of about 2. This variation reflects

**Table 2**

**Data from Kerem et al. (1989)**

| Probe | Enzyme | $n$ | $\hat{r}^2$ | $\hat{d}^2$ | $\hat{\Gamma}$ | CV ($\hat{d}^2$) |
|---|---|---|---|---|---|---|
| metD | *Ban*I | 160 | .1121 | .1127 | >10.0 | >1.9 |
| metD | *Taq*I | 172 | .0506 | .0236 | >10.0 | >1.9 |
| metH | *Taq*I | 152 | .0296 | .0282 | >10.0 | >1.9 |
| E6 | *Taq*I | 179 | .0468 | .0419 | >10.0 | >1.9 |
| E7 | *Taq*I | 164 | .0533 | .0486 | >10.0 | >1.9 |
| pH131 | *Hin*fI | 179 | .1759 | .1646 | 4.3 | 1.8 |
| W3D1.4 | *Hin*dIII | 184 | .1482 | .1413 | 6.0 | 1.9 |
| H2.3A | *Taq*I | 140 | .1093 | .0992 | >10.0 | >1.9 |
| EG1.4 | *Hin*cII | 163 | .3193 | .3042 | 1.3 | 1.4 |
| EG1.4 | *Bgl*II | 167 | .3441 | .3378 | 1.1 | 1.3 |
| JG2E1 | *Pst*I | 179 | .3280 | .3271 | 1.2 | 1.3 |
| E2.6 | *Msp*I | 121 | .2477 | .2193 | 2.7 | 1.6 |
| H2.8A | *Nco*I | 138 | .3186 | .3159 | 1.4 | 1.3 |
| E4.1 | *Msp*I | 147 | .1719 | .1461 | 5.9 | 1.8 |
| J44 | *Xba*I | 160 | .2297 | .1979 | 3.0 | 1.7 |
| 10-1X.6 | *Acc*I | 156 | .3938 | .3933 | .9 | 1.2 |
| 10-1X.6 | *Hae*III | 162 | .4101 | .4094 | .8 | 1.1 |
| T6/20 | *Msp*I | 151 | .0436 | .0271 | >10.0 | >1.9 |
| H1.3 | *Nco*I | 164 | .2790 | .2603 | 1.8 | 1.5 |
| CE1.0 | *Nde*I | 165 | .0101 | .0025 | >10.0 | >1.9 |
| J32 | *Sac*I | 130 | .0068 | .0061 | >10.0 | >1.9 |
| J3.11 | *Msp*I | 172 | .0214 | .0214 | >10.0 | >1.9 |
| J29 | *Pvu*II | 153 | .0331 | .0320 | >10.0 | >1.9 |

the stochastic nature of evolution, and it cannot be reduced with larger samples.

The estimates of $\Gamma$ in table 2 do not increase more or less monotonically as distance increases. This may be due to recombination hot spots, but we suspect that the large coefficients of variation and relatively small chromosomal region under consideration make it more likely that this behavior may be due to random error. The pessimism of Hill and Weir (1988) in using this approach to estimating $\Gamma$ may be justified, but a definitive answer will require additional data sets. If future data sets depart from expectation in different ways, then the explanation of random error will be strengthened; but, if they depart in a fashion consistent with that of table 2, then we will need to consider modifications to our model. Certainly the situation can be very complex, as shown by the recent paper by MacDonald et al. (1991) on linkage disequilibrium and Huntington disease.

All the results in the present paper are in terms of the conditional moments of $\hat{d}^2$, because they are easy to compute. Clearly it would be desirable to study the distribution of $\hat{d}^2$ values conditional on marker gene frequencies. The conditional probability that $\hat{d}^2$ exceeds some critical value would be of great interest, especially if it could be expressed as a function of $\Gamma$. At present it seems that conditional distributions of $\hat{d}^2$ will be found only from simulation, as in the work of Hudson (1985) on neutral genes. Hudson's approach can be used to simulate the conditional distribution of $\hat{d}^2$, providing that the $Q$ process is used to simulate the ancestral tree.

A priori, one does not know the allele frequency spectrum at the disease locus. If there is currently a predominant allele, as appears to be the case with CF, and if the frequency of this allele in the disease class has remained high for a long time, then the model can easily be modified, and the resulting calculations are not very different from the ones we present here. Alternatively, if the frequency of the predominant allele is drifting in time, or if there is no predominant allele but many low-frequency alleles, then the modifications needed to the model are not clear and are a subject of future research.

Finally, our analysis has considered markers one at a time. A more informative analysis should result from studying pairs of markers. In particular, it would be of interest to be able to infer whether a disease locus lies between two markers.

## Acknowledgments

Helpful comments on a draft version of this paper were received from Drs. J. Drake, B. Gladen, J. Taylor, and C. Weinberg. B.S.W. was supported in part by NIH grant GM45344.

## Appendix

### Identity Coefficients with Sampling

$$\phi^*(2,0) = \frac{1}{n_H} + \frac{n_H - 1}{n_H}\phi(2,0)$$

$$\phi^*(0,2) = \frac{1}{n_h} + \frac{n_h - 1}{n_h}\phi(0,2)$$

$$\phi^*(1,1) = \phi(1,1)$$

$$\phi^*(4,0) = \frac{1}{n_H^3} + \frac{7(n_H - 1)}{n_H^3}\phi(2,0) + \frac{6(n_H - 1)(n_H - 2)}{n_H^3}\phi(3,0) + \frac{(n_H - 1)(n_H - 2)(n_H - 3)}{n_H^3}\phi(4,0)$$

$$\phi^*(0,4) = \frac{1}{n_h^3} + \frac{7(n_h - 1)}{n_h^3}\phi(0,2) + \frac{6(n_h - 1)(n_h - 2)}{n_h^3}\phi(0,3) + \frac{(n_h - 1)(n_h - 2)(n_h - 3)}{n_h^3}\phi(0,4)$$

$$\phi^*(3,1) = \frac{1}{n_H^2}\phi(1,1) + \frac{3(n_H - 1)}{n_H^2}\phi(2,1) + \frac{(n_H - 1)(n_H - 2)}{n_H^2}\phi(3,1)$$

$$\phi^*(1,3) = \frac{1}{n_h^2}\phi(1,1) + \frac{3(n_h - 1)}{n_h^2}\phi(1,2) + \frac{(n_h - 1)(n_h - 2)}{n_h^2}\phi(1,3)$$

$$\phi^*(2,2) = \frac{1}{n_H n_h}\phi(1,1) + \frac{n_H - 1}{n_H n_h}\phi(2,1) + \frac{n_h - 1}{n_H n_h}\phi(1,2) + \frac{(n_H - 1)(n_h - 1)}{n_H n_h}\phi(2,2)$$

## References

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length of polymorphisms. Am J Hum Genet 32: 314–331

Chakraborty R (1986) Estimation of disequilibrium from conditional haplotype data: application to the β-globin gene cluster in American Blacks. Genet Epidemiol 3:323–333

Chakravarti A, Li CC, Buetow KH (1984) Estimation of the marker gene frequency and linkage disequilibrium from conditional marker data. Am J Hum Genet 36:177–186

Eiberg H, Mohr J, Schmiegelow K, Nielson LS, Williamson R (1985) Linkage relationships of paraoxonase (PON) with other markers: indication of PON-cystic fibrosis synteny. Clin Genet 28:265–271

Estivill X, Farrall M, Scambler PJ, Bell GM, Hawley KMF, Lench N, Bates GP, et al (1987) A candidate for the cystic fibrosis locus isolated by selection for methylation-free islands. Nature 326:840–845

Ewens WJ (1979) Mathematical population genetics. Springer, Berlin

Gusella JF (1991) Huntington's disease. In: Harris H, Hirschorn K (eds) Advances in genetics, vol 20. Plenum, New York, pp 125–151

Gusella JF, Wexler NS Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, et al (1983) A polymorphic marker genetically linked to Huntington's disease. Nature 306:234–238

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33:54–78

Hudson RR (1985) The sampling distribution of linkage disequilibrium under an infinite alleles model without selection. Genetics 109:611–631

Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. Genetics 120: 831–840

Kaplan NL, Darden T, Hudson RR (1988) The coalescent process in models with selection. Genetics 120:819–829

Kerem B-S, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. Genetics 61:893–903

——— (1971) Theoretical foundation of population genetics at the molecular level. Theor Popul Biol 2:174–208

Kupke KG, Graebert MB, Müller (1991) Mapping of the X-linked dystonia-parkinsonism syndrome (XDP) locus by analysis of linkage and linkage disequilibrium. Am J Hum Genet 49 [Suppl]: A347

MacDonald ME, Lin C, Srinidhi L, Bates G, Altherr M, Whaley WL, Lehrach H, et al (1991) Complex patterns of linkage disequilibrium in the Huntington disease region. Am J Hum Genet 49:723–734

Maiste PJ, Weir BS (1992) Estimating linkage disequilibrium from conditional data. Am J Hum Genet 50:1139–1140

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

Press WH, Flannery BP, Teukolsky SA, Vetterling WT

(1986) Numerical recipes: the art of scientific computing. Cambridge University Press, Cambridge

Pritchard C, Cox DR, Myers RM (1991) The end in sight for Huntington disease? Am J Hum Genet 49:1–6

Snell RG, Lazarou LP, Youngman S, Quarrell OWJ, Wasmuth JJ, Shaw DJ Harper PS (1989) Linkage disequilibrium in Huntington's disease: an improved localization for the gene. J Med Genet 26:673–675

Theilmann J, Kanani S, Shiang R, Robbins C, Quarrell O, Huggins M, Hedrick A, et al (1989) Non-random association between alleles detected at D4S95 and D4S98 and the Huntington's disease gene. J Med Genet 26:676–681

Tsui LC, Buchwald M (1991) Biochemical and molecular genetics of cystic fibrosis. In: Harris H, Hirschorn K (eds) Advances in genetics, vol 20. Plenum, New York, pp 153–266

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276

Weir BS (1989) Locating the cystic fibrosis gene on the basis of linkage disequilibrium with markers? In: Elston RC, Spence MA, Hodge SE, MacCluer JW (eds) Multipoint mapping and linkage based upon affected pedigree members: Genetic Analysis Workshop 6. Alan R Liss, New York, pp 81–86

Weir BS (1990) Genetic data analysis. Sinauer, Sunderland, MA

Weir BS, Hill WG (1980) Effect of mating structure on variation in linkage disequilibrium. Genetics 95:477–488

——— (1986) Nonuniform recombination within the human β-globin gene cluster. Am J Hum Genet 38:776–778