Am. J. Hum. Genet. 56:18–32, 1995

Likelihood Methods for Locating Disease Genes in Nonequilibrium Populations

N. L. Kaplan,[1] W. G. Hill,[2] and B. S. Weir[3]

[1]Statistics and Biomathematics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina; [2]Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh; and [3]Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh

Summary

Until recently, attempts to map disease genes on the basis of population associations with linked markers have been based on expected values of linkage disequilibrium. These methods suffer from the large variances imposed on disequilibrium measures by the evolutionary process, but a more serious problem for many diseases is that they assume an equilibrium population. For diseases that arose only a few hundred generations ago, it is more appropriate to concentrate on the initial growth phase of the disease. We invoke a Poisson branching process for this early growth, and estimate the likelihood for the recombination fraction between marker and disease loci, on the basis of simulated disease populations. The limits of the resulting support intervals for the recombination fraction vary inversely with the age of the disease in generations. We illustrate the procedure with data on cystic fibrosis and diastrophic dysplasia, for which the method appears appropriate, and for Friedreich ataxia and Huntington disease, for which it does not. A valuable aspect of the method is the ability in some cases to compare likelihoods of the three orders for a disease locus and two linked marker loci.

Introduction

For many rare genetic diseases, linkage analysis has been very successful in locating the disease gene. Linkage analysis requires recombinationally informative pedigrees, but the number of meioses observed in typical studies does not allow recombination events to be detected between loci within ~1 cM of each other (Hästbacka et al. 1992). More precise positioning of disease genes on a genetic map may be possible from population rather than family data, when measures of allelic association or linkage disequilibrium are estimated. This method depends on the comparison of marker allele frequency distributions in samples of normal and disease chromosomes. Differences among these distributions may imply linkage between disease and marker loci, since loci in a neighborhood of the disease locus on disease chromosomes are likely to carry the alleles present on the ancestral chromosomes on which the disease mutations first occurred. The size of the neighborhood will vary inversely with both the age of the disease mutation and the rate of recombination between marker and disease loci.

Although there are many evolutionary forces that can cause an association between alleles at different loci, only recombination is correlated with physical distance between the loci. For diseases that arose within the past few hundred generations, it is reasonable to assume that linkage disequilibrium reflects few recombination events between disease and marker loci. Marker alleles present on an ancestral chromosome in the neighborhood of a disease mutation will therefore be overrepresented in the disease population, and this is seen for example in data for cystic fibrosis (CF) (Kerem et al. 1989). In contrast, for diseases that have been in the population for thousands of generations, it is likely that disease chromosomes will carry many fewer ancestral marker alleles, because of the many more intervening recombination events, unless there is reduction of recombination in the region surrounding the disease locus. There can still be linkage disequilibrium between disease and nearby marker loci, as discussed by Hill and Weir (1994), but there will not be an excess of ancestral marker alleles on disease chromosomes. Assuming that there is no reduction in recombination, it follows that diseases for which linkage disequilibrium to several markers in a chromosomal region is observed, and for which the marker alleles are more frequent on disease than on normal chromosomes, are most likely of recent origin.

If the disease is not very old, and an association is detected between a marker and the disease, then it might be expected that at least one disease mutation has a high frequency in the population of disease chromosomes. Indeed, if all disease mutations are in low frequency, then, under the assumption that mutations occur at random, the ancestral disease haplotypes would be representative of those in the normal population, and there would be no subsequent association. CF, an autosomal recessive disorder, has one mutation responsible for ~70% of the disease

Received June 24, 1994; accepted for publication September 12, 1994.

Address for correspondence and reprints: Dr. Bruce S. Weir, Department of Statistics, North Carolina State University, Box 8203, 608 Cox Hall, Raleigh, NC 27695-8203.

0002-9297/95/5601-0004$02.00

population (Kerem et al. 1989), while Huntington disease (HD), a late-onset dominant disorder, may have one mutation accounting for about a third of the disease population (MacDonald et al. 1992). Also, in Finland, the majority of chromosomes carrying the diastrophic dysplasia mutation may have descended from a single mutation that was probably present in the founding population (Hästbacka et al. 1992).

A difficulty with population-association studies is quantifying the relation between recombination and degree of association. As recombination is not an observed event in these studies, the recombination fraction c must be estimated on the basis of a population genetic model. The models traditionally invoked assume that the human population has been of constant size and has reached an equilibrium under such forces as drift and recombination. Time is scaled in units of the effective population size, N_e, and equilibrium implies that the number of generations since the disease mutation is at least of the same order as N_e. Equilibrium models also have the recombination fraction confounded with N_e, although this quantity is generally unknown. In studies such as those of Chakravarti et al. (1984) or Estivill et al. (1987), the expected value of a statistic based on squared linkage disequilibrium was approximated by $1/(1 + 4N_ec)$, and N_ec was estimated by equating this quantity to the observed value of the statistic. A better procedure is to use maximum likelihood to estimate N_ec (Hill and Weir 1994). In addition to the confounding with N_e, there is the further problem that such estimates have very large variances because of the stochastic nature of the evolutionary forces that have shaped the population (Weir 1989; Weir and Hill 1986; Hill and Weir 1988, 1994). Empirical information on this genetic sampling would require information from replicate populations, which is unlikely to be possible. We have also previously noted (Weir and Hill 1980) that there is no simple algebraic expression (over all c values) for the equilibrium expected value of squared linkage disequilibrium and that account must be taken of the sampling framework of normal and disease genotypes (Kaplan and Weir 1992).

For a particular disease mutation, the probability that, on a disease chromosome randomly chosen from the current population, the chromosomal segment between disease and marker has remained intact since the time of the mutation is $(1 - c)^G \approx e^{-cG}$, if c is recombination fraction between the two loci and G is the number of generations since that time. Note that this equation confounds c with G instead of with N_e. Therefore, if linkage disequilibrium due to an excess of the ancestral marker allele in the sample of disease chromosomes is observed, we can conclude that e^{-cG} is high, meaning that cG is small. We might expect that $cG < 1$, since $e^{-1} = .37$ and >40% of the disease chromosomes are expected to carry the ancestral marker alleles. This conclusion has important implications. If, for example, G was as large as 5,000 generations (~100,000 years for 20-year generations), the condition would require $c < .0002$, meaning that every marker in linkage disequilibrium with the disease was fortuitously within 20 kb of the disease gene. (This assumes the usual rule of thumb of 1 cM $\approx$ 1,000 kb.) This is unlikely. In particular, in this scenario there is no basis for inferring ancestral haplotypes (Kerem et al. 1989; MacDonald et al. 1992), suggesting that a larger c and a smaller G is probably more appropriate. Hence, there is a need to focus on modeling the initial growth phase of the disease rather than its behavior under an equilibrium assumption.

The approach proposed by Hästbacka et al. (1992) for estimating cG was to equate e^{-cG} to the proportion of disease chromosomes in the sample carrying the inferred ancestral allele at a marker locus. In their data for diastrophic dysplasia in Finland, there were marker loci with high allele frequencies, so there was little doubt as to the ancestral type. The allele with highest frequency was logically assumed to be ancestral. This method has the same problem as the method based on equilibrium population theory, since the probability e^{-cG} refers to an expectation over all realizations of the evolutionary process. Equating observed and expected values of marker allele frequencies is, in effect, estimating an expected value by a single observation with attendant problems, unless the variance of this quantity is small.

As important as an estimate itself is a characterization of its sampling properties. In the present case, the important feature is the value of an upper confidence bound on the parameter c. Hästbacka et al. (1992) proposed confidence bounds on c, on the basis of Luria-Delbrück considerations, but the error distribution for these ad hoc bounds is not well characterized. As suggested earlier, the difficulty in studying the behavior of estimates of c is in taking into account the evolutionary history of the disease population, especially for equilibrium populations. An alternative to the Luria-Delbrück approach is to specify the evolutionary model and then simulate its dynamics. This was done for equilibrium models by Hill and Weir (1994), who used simulations to estimate the likelihood of N_ec for specific data sets.

In an upcoming paper, Kaplan and Weir (in press) modify the approach of Hill and Weir (1994), by assuming the disease is not old, and model its growth as a Poisson branching process. They estimate likelihoods of c for specified values G for the markers MET and D7S8 near the CF gene and obtain confidence bounds on c comparable to those obtained with linkage analysis. In this study we explore in greater detail the nonequilibrium approach of Kaplan and Weir. In particular, we generalize the method to accommodate the joint behavior of alleles at two marker loci with a view to determining the order among these loci and the disease locus. We compare the bounds on c found from this likelihood approach with those obtained by the method proposed by Hästbacka et al. (1992),

and we further illustrate the method with several additional data sets from the literature.

Methods

To make the presentation self-contained, we review the method of Kaplan and Weir (in press). A marker M is of interest because it is found to be in linkage disequilibrium with the disease. For simplicity, we assume it has two alleles, but all of the analyses can be extended to such multiple-allele markers as microsatellites. The most frequent allele in the disease sample is denoted M_1 and is assumed to be the marker allele on the ancestral disease chromosome. The other marker allele is written as M_2. For microsatellites it may not be clear which of several more or less equally frequent alleles in the disease sample is the ancestral allele. We suppose, further, that the disease is so rare that the marker polymorphism must have existed in the population at the time of the occurrence of the initial disease mutation. Even if several mutations have resulted in the same disease phenotype, the detection of linkage disequilibrium, between disease status and marker type, suggests that at least one of the disease mutations is found with high frequency in the disease population. This is the case with the ΔF_{508} mutation causing CF. For simplicity, then, we will assume, from now on, that there was a single disease mutation and that it occurred on a chromosome carrying marker allele M_1.

Because the disease is in low frequency, we assume that individuals are either heterozygous at the disease locus or homozygous for the normal allele. For relatively young diseases, the normal chromosome marker allele frequencies p_{1_n}, p_{2_n}, for alleles M_1, M_2, will be assumed constant over time. Within the disease population, however, the marker allele frequencies, p_{1_d} and p_{2_d}, are changing, and it is the stochastic process dictating these changes that we need to model. Sample sizes are written as k_n, k_d for normal and disease chromosomes.

Frequency changes between generations are governed by a Wright-Fisher sampling scheme. Although there may be a selective advantage for carriers, all carrier individuals are assumed to be selectively equivalent, as is reasonable to be assumed for a recessive disease. For a dominant disease, selective equivalence may also be reasonable to be assumed for such late-onset (postreproductive age) cases as HD. Time will be measured in generations, with $t = 0$ being the generation in which the disease mutation occurred, and $t = G$ being the generation from which the samples are taken. We adopt the simplification of nonoverlapping generations. In generation t there are $X_T(t)$ disease chromosomes, partitioned as $X_T(t) = X_1(t) + X_2(t)$, where $X_i(t)$ carry marker M_i, $i = 1, 2$.

Some disease chromosomes have not undergone recombination between disease and marker loci, since the time of the disease mutation. Those that have undergone recombination (necessarily with a normal chromosome, since disease homozygotes are assumed to be absent or not to contribute to the next generation), acquire marker allele M_i with frequency p_{i_n}. In the gamete pool from which generation $t + 1$ is formed, therefore, the fraction g_i of disease chromosomes carrying M_i is $g_i = \{1/2N\}[(1 - c)X_i(t) + cX_T(t)p_{i_n}]$, where N is the number of individuals in generation t. This size is not assumed to be constant.

As long as g_i is small, $X_i(t + 1)$, $i = 1, 2$, can be modeled as two independent Poisson random variables with means $2N(1 + \lambda)g_i$ (Ewens 1979). The quantity λ is small compared with 1 and can be interpreted as the sum of two quantities: ρ, the growth rate of the overall population, and s, the possible selective advantage of carriers over normals. Since the disease population has grown in size, $\lambda > 0$. The population size N cancels out of expressions for the mean, leading to the stochastic recursive relationships

$$X_i(t + 1) \sim \text{Poisson}\{(1 + \lambda)[(1 - c)X_i(t) + cX_T(t)p_{i_n}]\}, \qquad i = 1, 2, \quad (1)$$

where Poisson$\{Z\}$ denotes a Poisson variable with mean Z. Initially, $X_T(0) = X_1(0) = 1$, and $X_2(0) = 0$. If there are more than two marker alleles, equation (1) holds for each allele.

The stochastic recursion equation (1) can be used to simulate the evolution of the disease population for any set of values of the parameters λ, c, and p_{i_n}. As time progresses after the introduction of the disease allele into the population, the expected fraction of the disease population carrying marker allele M_i changes from its initial value to the normal population frequency p_{i_n}. The rate of change in p_{i_d} increases with recombination value c, and a given level of linkage disequilibrium corresponds to younger diseases as c increases. This inverse relationship between c and G is seen in the expectation of p_{1_d} (Cox et al. 1989):

$$\begin{aligned}\mathcal{E}(p_{1_d}) &= (1 - c)^G + [1 - (1 - c)^G]p_{1_n} \\ &\approx e^{-cG} + (1 - e^{-cG})p_{1_n},\end{aligned}$$

since c is small.

A reasonable range of values for c is from zero to a few cM. An obvious value to assign to p_{i_n} is the observed value f_{i_n} in the sample of k_n normal chromosomes, although binomial sampling can be invoked to give an approximate upper 95% confidence limit $f_{i_n} + 1.65\sqrt{f_{i_n}(1 - f_{i_n})/k_n}$. There is less guidance available for values of λ and G. We therefore try to choose a range of values for these two parameters in such a way that simulation results change very little over these ranges. This choice exploits information about the size of the current disease population.

The size $X_T(G)$ of the disease population contains information about λ and G, and the first step is to estimate this size. Often an estimate of the incidence of the disease in the population under study is available. Kerem et al. (1989), for example, give an incidence for CF of ~ 1 in 2,000, or .0005, in Caucasian populations of western European origin. From such estimates and from knowledge of whether the disease is dominant or recessive, the frequency of disease chromosomes in the population can be inferred. For the recessive CF, this frequency is estimated as $\sqrt{.0005} \approx .02$. The size $X_T(G)$ then follows from the total size of the population, which we take to be the total Caucasian population of ~ 500 million people, or 10^9 chromosomes. Hence, $X_T(G)$ would be ~ 20 million for CF. Alternatively, historical records may be available, as was the case in Finland, where Hästbacka et al. (1992) estimated the number of chromosomes carrying the mutation for diastrophic dysplasia to be 80,000.

It seems reasonable in any application to consider only those simulated evolutionary histories that lead to values of $X_T(G)$ close to the estimated value. For the Finnish example of diastrophic dysplasia, we might condition on histories resulting in $X_T(G)$ values in the interval (50,000–110,000). The actual size of such an interval is arbitrary and has little effect on final solutions. A strength of simulation is that this conditioning is very easy to implement. Demanding that $X_T(G)$ lies in a specified interval puts constraints on λ and G, since the growth of the disease population is being described by a Poisson branching process whose offspring distribution has mean of $1 + \lambda$. For a given value of G it is necessary to choose λ so that $X_T(G)$ will most often fall into the specified interval. A poor choice of λ will make the simulations very inefficient. The mean value of $X_T(G)$ is $(1 + \lambda)^G$, and the probability that the disease does not become extinct is $\sim 2\lambda$ (Ewens 1979). This suggests that λ can be estimated from

$$X_T(G) = \frac{(1 + \lambda)^G}{2\lambda}. \tag{2}$$

Determination of $X_T(G)$ is not very precise, and it is therefore important that resulting estimates of c are not overly dependent on this quantity. Since the behavior of a branching process is governed by the offspring distribution, and since the Poisson distribution is completely determined by its mean, $1 + \lambda$, changes in λ will not materially change the behavior of the X process, as long as $1 + \lambda$ does not change very much. The substantial change in $X_T(G)$ from 100,000 to 20,000,000 changes $1 + \lambda$ by >3% (λ changes from .047 to .078) when $G = 200$. For CF, for example, it is necessary to know only that the size of the disease population is in the millions, and conclusions about c do not change very much if that size is 1 million or 20 million.

It remains to assign a value to G. As discussed earlier, values of e^{-cG} that are not small imply that G is not large, since, otherwise, c would be too small. In the examples, we found that values of G in the hundreds gave reasonable results. We have also shown (Kaplan and Weir, in press) that underestimating G, which is the most likely error, is conservative because it leads to an overestimate of c. Of course, any external information about G should always be used, as when there are historical records (Hästbacka et al. 1992).

Once values for p_{i_n}, λ, and G have been decided, simulations are used to make inferences about c. For a given data set, the likelihood of c is estimated. This is easy to do, since, conditional on p_{1_d}, the number of disease chromosomes carrying M_1 has a binomial distribution with parameters k_d and p_{1_d}. The binomial probability, apart from the factorial terms, is $p_{1_d}^{n_1} p_{2_d}^{n_2}$, where the disease sample has $n_i = k_d f_{i_d}$ chromosomes with marker M_i. To estimate the likelihood, we simulate the population for G generations, always requiring that $X_T(G)$ be near its estimated value, to obtain p_{1_d}. We then average the associated binomial probabilities from repeated simulations. The only numerical difficulty is that the likelihood must be scaled in order to obtain non-negligible values, and we scale by dividing by $f_{1_d}^{n_1} f_{2_d}^{n_2}$. By repeating this process for many values of c, we establish the likelihood function. As is usual, we decrease the log-likelihood by two units from its maximum value, to establish a support interval for c corresponding approximately to a 95% confidence interval. To find the maximum, the likelihood is evaluated for values of c that were multiples of a small number: usually 0.001, but sometimes as small as 0.0002. For multiple-allele markers, we use a multinomial instead of a binomial distribution.

This method for generating confidence levels is heavily dependent on simulated data, in contrast to the method of Hästbacka et al. (1992), which used only experimental data. These authors appealed to Luria-Delbrück arguments to find confidence limits (c_-, c_+) satisfying

$$\begin{aligned} \frac{c_+}{\lambda}\left[\ln \frac{X_T(G)c_+}{\lambda} - 2\right] &= 1 - f_{1_d} \\ \frac{c_-}{\lambda}\left[\ln \frac{X_T(G)c_-}{\lambda} + 2\right] &= 1 - f_{1_d}. \end{aligned} \tag{3}$$

The performance of these bounds is examined in the Results section.

The model can be generalized to accommodate two marker loci. There is then the added feature of having to specify the relative positions of the disease and marker loci. The previous notation can be extended with an additional subscript for the second locus. For example, disease chromosomes with alleles i and j at the two marker loci have population frequency p_{ij_d}. The recombination frac-

tion between disease locus and the kth marker is c_k. Stochastic recursion equation (1) has three extensions, one for each of the three possible orderings of the loci. These equations are as follows, where a dot subscript indicates summation over all values of the corresponding index, and both c_1 and c_2 are assumed to be small. The equations assume no interference.

Marker-1–Disease–Marker-2

$$X_{ij}(t+1) \sim \text{Poisson}\{(1+\lambda)[(1-c_1-c_2)X_{ij}(t) + c_1 p_{i.n} X_{.j}(t) + c_2 p_{.jn} X_{i.}(t)]\}.$$

Marker-1–Marker-2–Disease

$$X_{ij}(t+1) \sim \text{Poisson}\{(1+\lambda)[(1-c_1)X_{ij}(t) + c_2 p_{ijn} \textstyle\sum_{k\neq i, l\neq j} \times X_{kl}(t) + (c_1-c_2)p_{i.n} X_{.j}(t)]\}.$$

Disease–Marker-1–Marker-2

$$X_{ij}(t+1) \sim \text{Poisson}\{(1+\lambda)[(1-c_2)X_{ij}(t) + c_1 p_{ijn} \textstyle\sum_{k\neq i, l\neq j} \times X_{kl}(t) + (c_2-c_1)p_{.jn} X_{i.}(t)]\}.$$

The parameters of the model are specified as before. For each order, and specified values of c_1, c_2, the p_{ij_d} values can be simulated, and observed frequencies can be compared with the corresponding multinomial samples. When the recombination value c_{12} between the markers is known, and the assumption of no interference is made, there is only one unknown recombination value.

The three different orderings of the two markers and disease locus can be compared on the basis of likelihoods. The haplotype counts in the disease sample have a multinomial distribution, with parameters given by the haplotype frequencies in the disease population, and the likelihoods can be estimated by simulation. Again, just as for a single marker, it is necessary to normalize the likelihoods by dividing by $\prod_i f_{id}^{n_i}$ where f_{id}, n_i, and $i = 1, 2, 3, 4$ are the sample frequencies and counts of the ith haplotypes in the disease sample.

Results

We illustrate the method by reference to four diseases, rather than by extensive simulations, partly because of the number of nuisance parameters that must be specified. The diseases are CF, diastrophic dysplasia, Friedreich ataxia (FA) and HD. We assumed that the most frequent marker allele on the disease chromosomes was the ancestral allele, i.e., the marker allele on the chromosome on which the disease mutation arose. In most cases this was an obvious choice, but for such highly variable markers as microsatellites the choice may not be as clear. For convenience, we always take the observed marker allele frequencies in the normal sample to be the population frequencies for the simulations, and we always used 1,000 simulations to estimate the likelihoods. We found that 1,000 gave a reliable picture of the likelihood curve in several test cases when we also used 10,000 simulations.

CF

CF is an autosomal recessive disease that provides the best example of the utility of linkage disequilibrium in mapping disease genes. As already noted, linkage disequilibrium for CF has been found in several Caucasian populations of western European origin. Therefore, we multiplied the frequency, .02, of the disease gene by 10^9 to obtain $X_T(G) = 2 \times 10^7$. Accordingly, we retain only those simulation runs for which the size of the disease population is in the interval (1.6×10^7, and 2.4×10^7). The size of the interval is arbitrary, and the main effect of this size is in determining the number of simulations needed. If the age G of the disease is specified, then the growth rate λ follows from $(1+\lambda)^G/2\lambda = 2 \times 10^7$.

Likelihoods were calculated with $G = 200$ and $\lambda = 0.078$. As Kaplan and Weir (in press) showed, the likelihoods are essentially the same for the same values of cG, so that support intervals for c, using other values of G can be obtained from those given here by multiplying them by $200/G$. In tables 1 and 2 we show estimates, and upper and lower support limits, for c for the markers MET, D7S8, XV2C, and KM19. For MET and D7S8, the likelihood was evaluated for values of c that were multiples of 0.001, whereas for XV2C and KM19 the values of c were multiples of 0.0005. We have used $G = 200$ in all calculations and have documented elsewhere (Kaplan et al., in press) our doubts about the recent revision in the age of the ΔF_{508} mutation for CF to 2,600 generations (Morral et al. 1994).

The first extensive sets of haplotype data for MET and D7S8 were published by Beaudet et al. (1986). The data were from several laboratories, and we have pooled them for this study. In table 1 we show bounds on c for the MET polymorphisms p*met*D/*Taq*I, p*met*H/*Taq*I, and p*met*H/*Msp*I. The p*met*D/*Taq*I allele frequencies from the London data were different from those from the other laboratories, so the bounds were calculated with those data either included or excluded. The data given by Beaudet et al. (1986) for p*met*H/*Taq*I did not show evidence of significant linkage disequilibrium and so were not used. However, two other data sets for this marker did have significant χ^2 statistics and so were included. All of the data sets except two, p*met*H/*Msp*I of Beaudet et al. (1986) and p*met*H/*Taq*I of Kerem et al. (1989), suggest that a conservative upper bound on c is .01. The two discrepant data sets suggest slightly larger bounds (.013 and .014). The lower bound on c is ~.002.

The χ^2 values in table 1 are those for the 2×2 contingency tables of disease versus marker types, and they indicate the strength of the association between the two loci. As expected, the likelihood curves in figure 1 show likelihood curves with less spread for higher χ^2 values.

Table 1

Upper and Lower Support Limits for the Recombination Fraction between the CF Gene and Markers MET and D7S8 (under the Assumption That $G = 200$ and $\lambda = .078$)

Marker	Greatest Marker-Allele Frequency		Disease Sample Size	χ^2	Lower Limit	Upper Limit	Maximum-Likelihood Estimate
	Disease	Normal					
p*met*D/*Taq*I[a]	.88	.77	402	17.4	.002	.008	.005
p*met*D/*Taq*I[b]	.92	.77	296	25.1	.001	.006	.003
p*met*D/*Taq*I[c]	.95	.80	79	8.7	.000	.006	.002
p*met*H/*Taq*I[d]	.73	.47	115	15.0	.001	.008	.004
p*met*H/*Taq*I[c]	.71	.54	69	4.5	.002	.014	.006
p*met*H/*Msp*I[a]	.71	.57	150	6.5	.003	.013	.007
p*met*H/p*met*D[a]	.62	.42	348	29.0	.003	.010	.005
p*met*H/p*met*D[d]	.64	.34	115	20.3	.003	.008	.005
pJ3:11/*Msp*I[a]	.56	.41	448	18.5	.005	.012	.008

[a] Beaudet et al. (1986).
[b] Beaudet et al. (1986) without London data.
[c] Kerem et al. (1989).
[d] Cutting et al. (1989).

Table 2

Upper and Lower Support Limits for the Recombination Fraction between the CF Gene and Markers XV2C and KM19 (under the Assumption That $G = 200$ and $\lambda = .078$)

Marker and Country	Greatest Marker-Allele Frequency		Disease Sample Size	Lower Limit	Upper Limit	Maximum-Likelihood Estimate
	Disease	Normal				
XV2C:						
Denmark[a]	.97	.38	280	.0001	.0009	.0004
Germany[a]	.86	.46	178	.0005	.0035	.0020
France[a]	.87	.52	680	.0010	.0035	.0020
Great Britain[a]	.88	.48	60	.0005	.0040	.0015
Spain[a]	.73	.50	100	.0020	.0080	.0050
Italy[b]	.74	.48	254	.0020	.0070	.0050
Northern Europe[c]	.83	.51	64	.0000	.0060	.0030
United States[d]	.92	.46	409	.0000	.0018	.0010
KM19:						
Denmark[a]	.94	.25	280	.0002	.0011	.0006
Germany[a]	.86	.31	178	.0006	.0024	.0016
France[a]	.90	.28	680	.0004	.0015	.0010
Great Britain[a]	.90	.27	60	.0000	.0024	.0010
Spain[a]	.72	.34	100	.0010	.0060	.0030
Italy[b]	.79	.30	228	.0010	.0035	.0025
Northern Europe[c]	.88	.30	80	.0000	.0025	.0015
United States[d]	.93	.27	409	.0004	.0014	.0010

[a] Serre et al. (1990).
[b] Estivill et al. (1988).
[c] Kerem et al. (1989).
[d] Cutting et al. (1989) plus Beaudet et al. (1989).

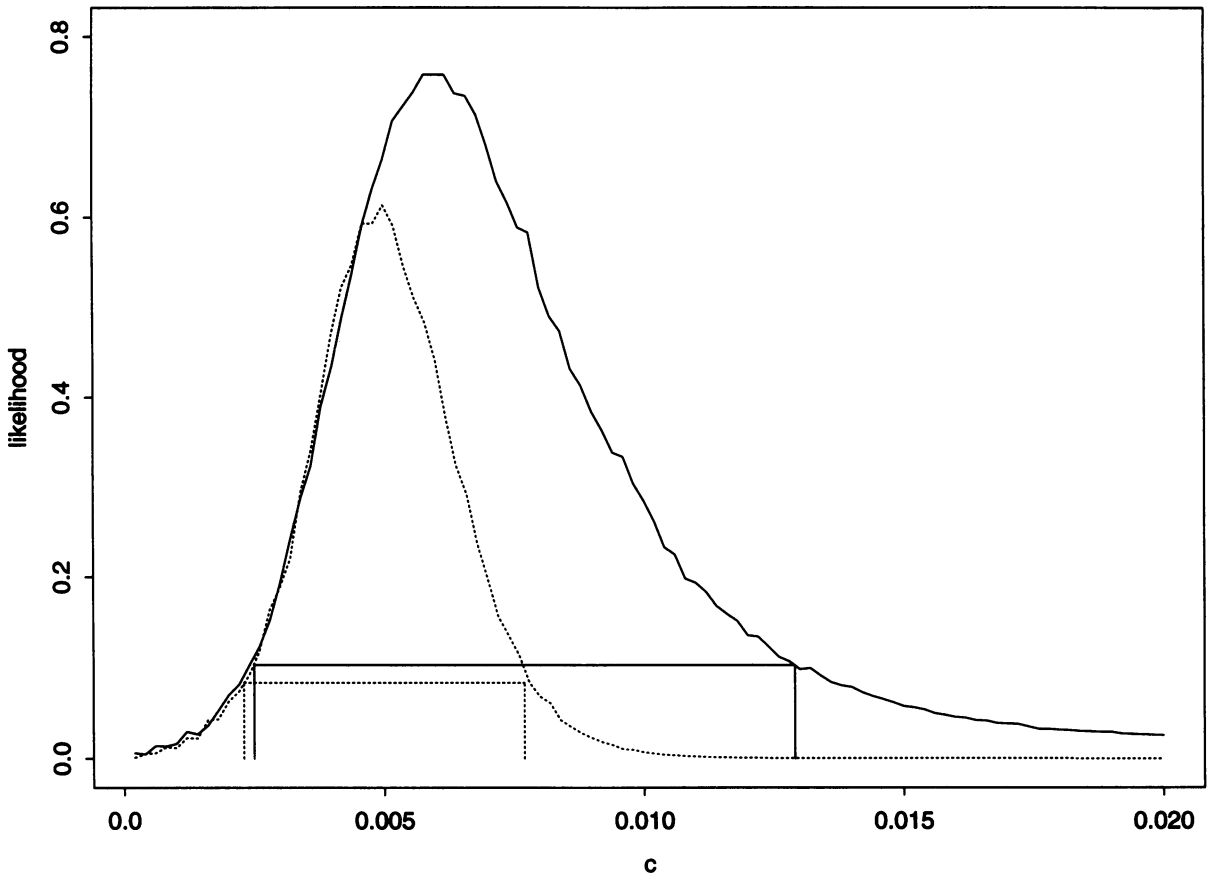

Figure 1 The likelihood function for two samples from table 1, showing the effects of the strength of the association between disease and marker loci. The dotted line is for p*met*D/*Taq*I[a] data with a χ^2 of 17.4, and the solid line is for p*met*H/*Taq*I[c] with a χ^2 of 4.5. The 2-LOD intervals are also indicated.

Since p*met*D/*Taq*I and p*met*H/*Taq*I are so close, we also considered them as constituting a single three-allele marker (see Beaudet et al. 1986, table 7). We excluded the very rare double recombinant, and the results also support the bounds of .002 and .01 on *c*. Table 1 also shows bounds on *c* for the flanking marker pJ3.11=D7S8. These results show that upper and lower bounds on *c* for the polymorphism pJ3.11/*Msp*I are still .012 and .005, for $G = 200$.

Beaudet et al. (1986) used linkage analysis to obtain upper confidence bounds of .012 and .011 from the disease locus to MET and D7S8, respectively. These bounds are consistent with those given in table 1. The lower bounds in the table are >0, whereas Beaudet et al. could not exclude zero recombination rates. Kerem et al. (1989) gave a physical map for the region surrounding the CF gene, showing ~900 kb from MET to CF and ~800 kb from D7S8 to CF. These translate to recombination fractions of .009 and .008, if 1 cM = 1,000 kb.

In table 2 support limits on *c* are given for two closer markers, XV2C and KM19. Data and results are shown for six European countries and two pooled samples, northern Europe and United States. Apart from Denmark, which may have been more isolated than the other European countries sampled, the most frequent marker alleles in the normal populations had very similar frequencies. However, the most frequent marker alleles in the CF populations show a clear increase from south to north. Spain and Italy have lower frequencies than do Germany, France, and Great Britain, while Denmark has the highest frequencies. The simplest explanation is that the CF mutation spread across Europe by migration. The higher the frequency of the most frequent marker allele in the disease population, the more recent was the introduction of the disease into the population, and the less time there has been for the association between disease and marker to have been weakened by recombination. This hypothesis has been discussed in detail by Serre et al. (1990). The calculations in table 2 use 2×10^7 for the worldwide size of the CF population. We repeated the calculations, using the sizes for individual countries, and found little change in the results.

The data from France, Great Britain, and Germany suggest that the recombination fraction between XV2C and CF < .004, assuming $G \geq 200$. Spain and Italy give compa-

rable estimates if the disease is ⩾300 generations old in those countries. The bound for Denmark is much lower, suggesting the reasonable scenario of a more recent introduction of CF into Denmark. The same pattern is found for KM19, except that the bounds are somewhat smaller, which would suggest KM19 is closer to CF than is XV2C. The U.S. data lead to smaller bounds on c for XV2C and KM19 than any of the European data, with the exception of Denmark.

There do not appear to be any linkage data between CF and XV2C or KM19, probably because these markers are so close to the CF gene. Values have been suggested, from indirect evidence, for the recombination values from CF to these markers by Farrall et al. (1988) and Beaudet et al. (1989).

Under the assumption that the CF mutation is ⩾200 generations old, the bounds in table 2 for Germany, France, and Great Britain are consistent with the physical distances given by Kerem et al. (1989). These authors' estimates of c for XV2C and KM19 are .003 and .023, respectively. Serre et al. (1990) set up transition equations for marker haplotype frequencies and manipulated these to arrive at a moment estimator for the recombination fraction from French data. They reported .0015 for XV2C and .0008 for KM19, assuming $G = 200$. These values are lower than those in table 2, probably because of the use of expected values by Serre et al. (1990). The moment estimate of Hästbacka et al. (1992) also gives underestimates for the recombination fraction c.

Two-locus data were considered for determining the order between MET, D7S8 and CF. Using linkage analysis, Beaudet et al. (1986) estimated c_{12}, the recombination fraction between MET and D7S8, to be .018, with a confidence region of .008–.036. For several different values of c_{12}, we maximized the likelihood for each of the three possible orderings of p*met*H/*Taq*I, p*J311*/*Msp*I, and CF, by using the haplotype data given by Beaudet et al. (1986). We assumed that $G = 200$ and that $\lambda = .078$. For CF between the two markers, the likelihood function was evaluated at nine equally spaced points along the marker interval, and the maximum value was plotted in figure 2. For the other two orders, the likelihood was evaluated for a series of recombination values from CF to the nearest marker calculated as multiples of 0.001, and the maximum of these values was plotted in figure 2. Clearly, the order with best support has CF between the two markers. For $c_{12} = .01$, however, the odds against the order CF-MET-D7S8 are only 5:1. Because the curve for this order decreases very rapidly as c_{12} increases, and because .01 is at the lower end of the support interval for c_{12}, the data do not lend strong support to the order. In almost all cases, the likelihood for the MET–CF–D7S8 order was maximized for $c_1 = 3c_{12}/8$, so, for example, if $c_{12} = .018$, the estimates of c_1 (CF–MET) and c_2 (CF–D7S8) are .007 and .011, respectively. When the likelihood calculations were

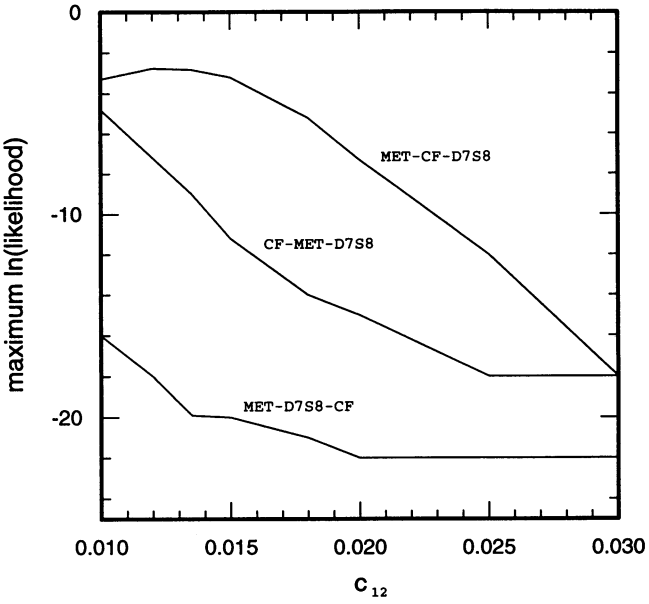

Figure 2 For each of the three orderings of CF, MET, and D7S8, the maximum value of the natural logarithm of the likelihood is plotted as a function of c_{12}, the recombination fraction between MET and D7S8.

repeated with $G = 300$, $c_{12} > .015$, the curves were significantly lower than those for $G = 200$, suggesting that $c_{12} > .015$ and $G = 300$ are not compatible. Recently, Serre et al. (1990) also concluded that G lies between 150 and 250.

Since the markers XV2C and KM19 are so close, we ignored the possibility that the CF gene lies between them, and considered only the other two orderings. We evaluated likelihoods for the two values of the recombination fraction c_{12} between these two markers: .0006 (Kerem et al. 1989) and .001 (Beaudet et al. 1989). We used $G = 200$ and $\lambda = .078$, and chose c_1, c_2 values to maximize likelihoods for c_{12} values that were multiples of 0.0004. Data from France (Serre et al. 1990), Italy (Estivill et al. 1988), and the United States (Cutting et al. 1989) were analyzed. The results in table 3 strongly support the order XV2C–KM19–CF.

Diastrophic Dysplasia

Diastrophic dysplasia is an autosomal recessive disease. Hästbacka et al. (1990) used linkage analysis to localize, to chromosome 5, the gene (DTD) responsible for the disease, and Hästbacka et al. (1992) used linkage disequilibrium in Finnish data to give a more precise location. These Finnish data provide an excellent example for the present method. There is evidence that the current Finnish population descended from a founding population ~2,000 years ago and has subsequently been isolated. It is reasonable to hypothesize that the current frequency of DTD in Finland, ~0.8%, is the result of a founder event due to a population bottleneck. The present-day population size of

Table 3

Likelihoods for the Orders of XV2C, KM19, and CF (under the Assumption that G = 200 and λ = .078)

Sample and Order	c_{12}	Maximum ln(Likelihood)	Maximum-Likelihood Estimate of c[a]
France[b]:			
XV2C–KM19–CF	.0006	−2.54	.0012
CF–XV2C–KM19		−8.30	.0008
XV2C–KM19–CF	.001	−1.54	.0012
CF–XV2C–KM19		−9.80	.0008
Italy[c]:			
XV2C–KM19–CF	.0006	−2.42	.0020
CF–XV2C–KM19		−4.67	.0016
XV2C–KM19–CF	.001	−1.52	.0020
CF–XV2C–KM19		−5.45	.0016
North America[d]:			
XV2C–KM19–CF	.0006	−4.18	.0008
CF–XV2C–KM19		−5.68	.0004
XV2C–KM19–CF	.001	−3.60	.0008
CF–XV2C–KM19		−6.25	.0004

[a] Recombination fraction from CF to nearest marker.
[b] Serre et al. (1990).
[c] Estivill et al. (1988).
[d] Cutting et al. (1989).

Finland is ~5 million, so the disease population size is ~80,000 chromosomes. With G = 100 and 20 years per generation, the growth rate is λ = .102 if $X_T(0) = 1$. If $X_T(0) \neq 1$, equation (2) needs to be modified. Since the probability that the disease population does not go extinct is $\sim 1 - e^{-2X_T(0)\lambda}$, it follows from equation (2) that $X_T(G) = [X_T(0)(1 + \lambda)^G]/[1 - e^{-2X_T(0)\lambda}]$. Under the assumption of Hästbacka et al. (1992) that the founding population had a size of $\sim 10^3$, ~8 of these people are expected to have carried the disease mutation, and the corresponding value of λ is .094. The size of $X_T(0)$ has little effect on λ.

Hästbacka et al. (1992) used a slightly different approach for determining λ. They assumed a founding population size of 1,000 and a current population size of 5 million and related the two by exponential growth, i.e., $10^7 = e^{\lambda G}10^3$. This provides λ = .092, but it makes little difference to the results if we use λ = .092, .094, or .102, and we choose to use the last value.

Five informative polymorphisms were identified in the CSF1R gene on chromosome 5. The upper bounds on recombination fractions between the DTD locus and *Sty*I and *Eco*RI are .0026 and .0022, respectively, suggesting that the DTD gene is within ~250 kb of these polymorphisms. The bound given by Hästbacka et al. (1992) was .00086 (86 kb). The two-marker analysis was not informative, because the markers are too close together and because only one chromosome in the disease sample (the haplotype 1-2: allele 1 at locus *Sty*I and allele 2 at locus *Eco*RI) could have been a recombinant. There is no compelling evidence from linkage disequilibrium in favor of either of the orders DTD–*Sty*I–*Eco*RI or *Sty*I–*Eco*RI–DTD. We also treated *Sty*I–*Eco*RI as a single polymorphism with four alleles, and found an upper bound of .0021 for c.

Hästbacka et al. (1992) also typed three microsatellite markers in the neighborhood of the two RFLPs. Although the high mutation rate for these markers limits their use in association studies, they do provide useful evolutionary information. For example, the five-marker haplotype (CCTT–CA–*Sty*I–TAGA–*Eco*RI) for the single 1-2 *Sty–Eco* haplotype is 1-1-1-1-2, which is the same as the most frequent disease haplotype 1-1-1-1-1, apart from the *Eco*RI allele. This haplotype is not seen in the normal sample, suggesting that it has a low frequency in the normal population. The simplest explanation for the 1-1-1-1-2 haplotype is that a recombination event occurred between microsatellite TAGA and *Eco*RI, involving a disease chromosome with haplotype 1-1-1-1-1 and a normal chromosome having the *Eco*RI-2 allele. As Hästbacka et al. (1992) pointed out, this scenario would imply that the DTD gene is distal to the CSF1R gene.

The five-marker haplotypes of the eight DTD chromosomes whose *Sty*I–*Eco*RI haplotype differs from 1-1 call into question the original assumption that all these haplotypes are products of recombination. From table 2 of Hästbacka et al. (1992), seven of the eight had *Sty*I–*Eco*RI haplotype 2-2, six had *Sty*I–TAGA–*Eco*RI haplotype 2-5-2, five had CA–*Sty*I–TAGA–*Eco*RI haplotype 7-2-5-2, and

four had CCTT-CA-*Sty*I-TAGA-*Eco*RI haplotype 1-7-2-5-2. These haplotypes are such that at least four of them may be the consequence of a single recombination event. How could this have happened? With only recombination as a mechanism, a substantial fraction of the non-1-1-disease *Sty*I-*Eco*RI haplotypes are a consequence of a single recombination, and at least four (and possibly seven) of the eight were chosen at random from descendants of this event. This seems unlikely. Alternatively, there may have been additional disease mutations, or immigration of different disease haplotypes. (A reviewer pointed out that Finland received a large number of immigrants from Sweden during the Middle Ages [de la Chappelle 1993].) Mutation seems improbable, given that the 1-7-2-5-2 haplotype is rare in the normal population also. Even though the assumption that the eight non-1-1 *Sty*I-*Eco*RI disease haplotypes are all recombinants may not be correct, the assumption is conservative and, at worst, leads to an overestimate of the recombination fraction c.

Hästbacka et al. (1992) proposed using equation (3) to provide an upper bound on the recombination fraction. We performed simulations to evaluate the behavior of this bound. For each of four different values of c, .0006, .0012, .0018, and .0024, we simulated 100 samples of DTD chromosomes. Each sample was obtained by first simulating a population of DTD chromosomes and then taking a random sample of size 152. The simulation parameters were the same as those used to obtain the likelihood estimates of c for *Sty*I and *Eco*RI ($G = 100$, $\lambda = .102$, and $p_{1_n} = .26$). We found for small values of c that evolutionary forces skew the values of p_{1_d} toward 1. The estimate of Hästbacka et al. does not adequately compensate for this, so they underestimate the actual value <70% of the time in the simulations. We also found that the upper bound proposed by Hästbacka et al. was less than the actual value of $c \geq 40\%$ of the time. Moreover, those values of the bound less than the true value had a mean of about two-thirds the true value. The likelihood-based bound on c was almost always greater than the true value. Although this is satisfactory, the bound was typically at least twice the true value and so quite conservative (table 4).

FA

FA is an autosomal recessive neurodegenerative disease with an incidence reported to be 1 in 50,000 for the United Kingdom (Chamberlain et al. 1988) and 1 in 25,000 for Italy (Romeo et al. 1983). The FA gene, FRDA, was localized to chromosome 9, by linkage analysis (Chamberlain et al. 1988) and found to be within 1 cM of the tightly linked markers D9S5 and D9S15 (Chamberlain et al. 1993), with few recombinants observed between the disease and these markers (Rodius et al. 1994). Linkage disequilibrium between FA and the marker D9S15/D9S5 has been found in some populations (e.g., see Fujita et al. 1990; Hanauer et al. 1990; Pandolfo et al. 1990) but not in others (Chamberlain et al. 1993), suggesting either a founder effect (Sirugo et al. 1992) or multiple disease mutations, with none in high frequency. Much of the linkage disequilibrium data is from France (Fujita et al. 1990; Hanauer et al. 1990), so we consider only these data and use the size of the French population of approximately 10^8 chromosomes. Accordingly, $X_T(G) = 5 \times 10^5$.

Since there is no evidence for the age of the disease mutation, we used $G = 200$ and $\lambda = .056$ for the likelihood calculations. The resulting bound on c, from the D9S15/*Msp*I data of Fujita et al. (1990) is .012, which is comparable to the bound found by linkage analysis (Hanauer et al. 1990). If a smaller G were chosen, the bound would have been even larger. These data, therefore, do not provide much additional information regarding the location of the FA gene. For these data to be of help, the disease mutation would have to be older than 200 generations. The lower bound on c was .002, whereas Hanauer et al. (1990) could not rule out $c = 0$ with linkage analysis.

Fujita et al. (1990) presented additional polymorphisms at D9S5 and D9S15. The three-allele marker D9S5/26P-*Bst*XI data do exhibit linkage disequilibrium with FA, unlike the D9S5/*Msp*I polymorphism, although we found the resulting bound on c to be >.03, which is larger than the .015 value found by linkage analysis. They also identified a six-allele microsatellite marker for D9S15 that shows linkage disequilibrium with FA, and the likelihood bound on c is .02. The upper bounds on c were too large for these markers to be of great use. For each marker, the large bound results from both the low frequency of the most frequent marker allele and the small size of the disease sample. For example, the disease sample size for D9S15/*Msp*I was 76, and the highest marker allele frequency was .55, as compared with .3 in the normal sample. Apart from factors such as multiple disease mutations, gene conversion, and double recombination events, there is the additional possibility of mutation at this microsatellite marker, lowering the highest marker frequency in the disease population.

Pandolfo et al. (1990) sampled Italian families but did not find linkage disequilibrium between D9S5/26P-*Bst*XI and FA. This may be a sampling phenomenon. They did find linkage disequilibrium for the D9S15/*Msp*I polymorphism, but the M_1 allele had a frequency of .60 in the disease sample and of .80 in the normal sample. This situation is one in which the marker shows linkage disequilibrium with the disease, but the marker allele frequencies are not consistent with the theory in this paper. It is difficult to distinguish between noncompliance with the model and the effects of small sample size.

Additional evidence for possible ambiguities in linkage disequilibrium data for FA is provided by Chamberlain et al. (1993). No associations were found between disease and alleles at either D9S5 or D9S15 in a worldwide sample of 104 FA families or in a subset of 41 U.K. FA families.

Table 4

Comparison of Bounds on c Given by Hästbacka et al. (1992) and by Likelihood Method, by Using StyI Data of Hästbacka et al., on the Basis of 100 Simulations

	HÄSTBACKA ET AL.		LIKELIHOOD	
c	Mean (SD) Bound[a]	Proportion of Bounds < c	Mean (SD) Bounds[a]	Proportion of Bounds < c
.6	.36 (.11)	.49	1.82 (.67)	.00
1.2	.83 (.21)	.40	3.40 (1.30)	.01
1.8	1.33 (.31)	.66	3.95 (1.15)	.01
2.4	1.71 (.39)	.64	4.68 (1.11)	.01

NOTE.—Values of c and bounds are $\times 10^3$.

[a] Mean and SD are calculated conditional on the bound being less than c (Hästbacka et al.) or on the bound being greater than c (likelihood method).

The linkage studies of Chamberlain et al. (1993), however, did provide additional evidence that both D9S5 and D9S15 are within 1 cM of the disease locus. On the basis of a rare recombinant event, Chamberlain et al. (1993) proposed the order FA–D9S5–D9S15. If there were a single disease haplotype, this order could not support the extended haplotype data of Fujita et al. (1990), on the basis of only single recombination events. Given the rarity of double recombinants within map distances of 1 cM, it seems more likely, as suggested by Chamberlain et al. (1993), that there are multiple ancestral mutations, none of which have a high frequency in the disease population.

HD

HD is a very rare dominant, late-onset disease that affects ~1 in every 10,000 people of European descent (The Huntington's Disease Collaborative Research Group, 1993). The gene responsible for the disease has been localized by linkage analysis to a region of chromosome 4 (Gilliam et al. 1987; Whaley et al. 1988). Crossover events predicted two physically separated candidate regions containing the disease gene: a 100-kb terminal region near the telomere and a 2.5-Mb internal region between markers D4S10 and D4S168 (Andrew et al. 1992). Linkage disequilibrium data did not clarify the picture, although the evidence appeared to favor the internal region. Unlike the situation for CF, there is no predictable pattern to the marker associations, and so these data have not provided a more precise placement of the HD gene. There has also been discussion about possible misdiagnoses of the disease (Andrew et al. 1994).

An alternative strategy is to compare detailed haplotypes of HD and normal chromosomes in the candidate region and look for subregions containing a consistent HD haplotype. MacDonald et al. (1992) used a number of multiallele polymorphisms to study the internal candidate region and found that >40% of the HD chromosomes in their sample had one of two distinct haplotypes in a 500-kb subregion between markers D4S180 and D4S182. This suggests that these HD chromosomes may be ancestrally related and that the genetic defect may be in this subregion. Subsequent work by the Huntington Disease Collaborative Research Group (1993) identified in the subregion a new gene, IT15, with an expandable unstable trinucleotide repeat, which is believed to be the genetic alteration responsible for the disease.

The subregion identified by MacDonald et al. (1992) contains the marker D4S95, which shows consistent linkage disequilibrium with HD, for the polymorphisms *Acc*I and *Mbo*I (Thielmann et al. 1989; Adam et al. 1991; MacDonald et al. 1991; Andrew et al. 1992; Skraastad et al. 1992; Snell et al. 1992) but does not show disequilibrium for the nearby *Taq*I polymorphism. As pointed out by MacDonald et al. (1992), one explanation of this disequilibrium inconsistency is that the two most common haplotypes in their sample differ for *Taq*I alleles but not for *Acc*I or (presumably) *Mbo*I alleles. We proceeded with the likelihood analysis for the *Acc*I and *Mbo*I data, recognizing that there are multiple ancestral haplotypes and that the bounds may underestimate the true value of c.

For table 5, five data sets are considered in which there is evidence of linkage disequilibrium between HD and D4S95/*Acc*I. Two of these data sets are from the United Kingdom (Snell et al. 1989; Adam et al. 1991); one is from the Netherlands (Skraastad et al. 1992); one is from Canada, which can be regarded as a western European sample (Andrew et al. 1992); and one is a western European sample (MacDonald et al. 1991). We adopted the same approach as for CF and took the population size to be 10^9 chromosomes. The disease-population size was assumed to be 10^5, and we used $G = 200$ and $\lambda = 0.047$. The upper bounds on c in table 5 range from .008 to .015. We also pooled the data, to obtain a bound of .008. These values are in line with the conclusion of MacDonald et al. (1992), giving a 700-kb region around D4S95 as most probably containing the disease gene.

Table 5

Upper and Lower Support Limits for the Recombination Fraction between the HD Gene and Markers D4S95 and D4S127 (under the Assumption That $G = 200$ and $\lambda = .047$)

Data Set	Greatest Marker-Allele Frequency		Disease Sample Size	χ^2	Lower Limit	Upper Limit	Maximum-Likelihood Estimate
	Disease	Normal					
D4S95/*Acc*I:							
Snell et al. (1989)	.88	.59	41	11.1	.000	.008	.003
Adam et al. (1991)	.84	.64	56	8.0	.001	.012	.004
Skraastad et al. (1992)	.92	.68	24	5.2	.000	.011	.002
MacDonald et al. (1991)	.88	.73	51	4.6	.000	.015	.004
Andrew et al. (1992)	.81	.68	112	8.1	.001	.015	.006
Pooled data	.85	.67	284	41.9	.001	.009	.005
D4S95/*Mbo*I:							
Snell et al. (1989)	.90	.63	51	11.9	.000	.007	.002
Adam et al. (1991)	.84	.56	51	13.5	.001	.008	.003
Skraastad et al. (1992)	.89	.57	45	14.9	.000	.006	.002
Andrew et al. (1992)	.81	.64	115	12.2	.001	.012	.005
Pooled data	.85	.61	262	50.1	.001	.007	.004
D4S127/*Pvu*II:							
MacDonald et al. (1991)	.84	.58	58	12.7	.000	.008	.003

Also included in table 5 are results for four data sets, two U.K., one Netherlands, and one Canadian, showing linkage disequilibrium for the D4S95/*Mbo*I polymorphism. The bounds in these cases are slightly smaller, with the pooled data suggesting a bound of .007. The final data set in table 5 is for the marker D4S127, located near D4S95 and included in the haplotype analysis of MacDonald et al. (1992). The bound on c was found to be .008.

The marker D4S98 also shows linkage disequilibrium with HD (Snell et al. 1989; Thielmann et al. 1989; MacDonald et al. 1992), but the allele frequency pattern at this locus is not consistent with our evolutionary model. The most frequent marker allele among the HD chromosomes is less frequent than it is among the normal chromosomes. The χ^2 statistic for association between marker and disease is large for reasons other than discussed here, and we caution against the use of this marker. Either a small sample size or population heterogeneity may be responsible for this finding, and indeed, linkage disequilibrium for D4S98 and HD is not found in the data of either Skraastad et al. (1992) or Andrew et al. (1992).

There are many reasons to be cautious when interpreting linkage disequilibrium results for HD. Andrew et al. (1992) showed three tightly linked markers ~3 Mb telomeric to D4S95 that were in linkage disequilibrium with HD, although several intervening markers did not show allelic associations with the disease. This behavior is not expected if it is recombination that is responsible for marker variation in the disease population. One obvious explanation for this unusual association pattern is the occurrence of multiple ancestral mutations at different locations on chromosome 4. MacDonald et al. (1992) haplotyped 78 HD chromosomes in the region around D4S95, identified two possible ancestral haplotypes comprising about one-third of the sample, and suggested that there may be more. It would be very interesting if the chromosomes were also typed for the markers identified by Andrew et al. (1992), to see whether haplotype patterns can be explained by the multiple mutation hypothesis.

Discussion

Linkage analysis has proved to be very successful in localizing disease genes. Its usefulness is limited in fine-scale mapping, however, because of the difficulty in obtaining recombinationally informative families. An alternative strategy based on linkage disequilibrium has been used with some success. This approach, which is population based rather than family based, compares marker allele frequencies in normal and disease chromosomes. If these frequency distributions are judged to be different, then the marker is of interest and may be close to the disease locus.

It is not easy to quantify physical or genetic distance on the basis of linkage disequilibrium. We have criticized previous attempts to estimate recombination values c between disease and marker loci from linkage disequilibria on several grounds, including the large variances associated with such estimates and lack of attention to sampling strategies. The issue we have taken up here concerns the underlying evolutionary model. Until the work of Hästbacka et al. (1992), previous attempts to estimate c have assumed a state of equilibrium between the forces of drift and recombination, and sometimes mutation. Equilibrium models are almost certainly not appropriate for rare human diseases.

Linkage disequilibrium analysis rests on the simple observation that descendants of a disease mutation tend to share the ancestral haplotype in the neighborhood of the disease locus. The size of this neighborhood depends on the recombination rate c and the age G of the mutation, with the product cG being the critical parameter. The value of this parameter cannot be large if the current population contains ancestral disease haplotypes. The equilibrium model requires large G values, and hence values of c that are unrealistically small for all markers in linkage disequilibrium with the disease. This problem is overcome by abandoning the equilibrium model, as must be done if ancestral haplotypes are present in the current disease population. This conclusion holds in general and not just for isolated populations, as might be inferred from the Finnish study of Hästbacka et al. (1992).

Under reasonable assumptions, a Poisson branching process gives an adequate description of the early growth of the disease population. Simulation can be used to provide maximum likelihood bounds for the recombination fraction c between a marker and the disease locus. The evolution of marker allele frequencies in the disease population is simulated many times, and samples are drawn from these simulated populations for the likelihood calculations. An advantage of basing estimation on simulations is that additional information can be incorporated into the analysis. For example, if there is information on the size of the disease population, then it can be required that simulated disease populations have a size close to this value. The results are fairly robust to the several parameters that must be specified for simulation, except for the age G of the disease. Fortunately, G acts as a scale parameter (Kaplan and Weir, in press), so a single value can be used in simulations. This still allows the comparison of results over several marker loci, for which G will be the same.

To carry out the simulations it is necessary to specify the ancestral marker allele. Usually, the obvious choice is the most frequent allele in the disease sample. In cases where the choice is ambiguous, such as for microsatellite markers, computing the likelihood conditional on each allele being ancestral, weighting each likelihood by the a priori probability on ancestry and then maximizing the overall likelihood seems to be a reasonable strategy that would merit further research.

Another benefit of using simulations is that it is not necessary to assume a constant population size. For the branching process model, changing the population size causes λ to change, and it is a simple matter to incorporate these changes into the simulation. The human population has increased rapidly, especially in recent generations, and one simple way that this may be modeled is with a constant λ for all but the most recent generations. Since allele frequencies do not change very much in large populations, we would not expect this modification to alter our conclusions very much. This was borne out by simulation.

An idea of the success of the simulation approach is provided by an examination of the marker allele distribution in the disease sample. The theory adopted here supposes one marker allele in the initial disease population, and then a decrease in the frequency of this allele to the value in the general population. Detection of linkage disequilibrium, where the most frequent marker allele has a lower frequency in the disease than in the normal population, indicates that the theory does not hold in those cases: e.g., D4S98/*Sst*I and HD (Snell et al. 1989) and D9S15/*Msp*I and FA (Pandolfo et al. 1990). Caution should be exercised for markers with this kind of allele frequency distribution.

Four human diseases were discussed in this paper. Recombination values found by the likelihood approach for CF were consistent with the physical map given by Kerem et al. (1989). It is instructive to point out why CF is the best example of the use of linkage disequilibrium to map a gene. In the first place, most disease chromosomes descend from a single ancestral mutation, ΔF_{508}. Second, most of the markers on the ancestral haplotype in the vicinity of the disease locus had alleles with low frequencies in the general population. Finally, the mutation is recent enough so that recombination has not yet had the opportunity to break up the ancestral haplotype. This argument was made by Serre et al. (1990) in arguing for the disease being young. (An alternative hypothesis, based entirely on arguments about mutation, was made by Morral et al. [1994] to argue for a much greater age. We have doubts about this argument [Kaplan et al., in press].) This evolutionary scenario made CF a lucky" case for using linkage disequilibrium, and there is no reason to expect other diseases to have such a favorable history.

Indeed, linkage disequilibrium was not very informative for FA or HD. For FA, the most frequent marker alleles in the disease population have relatively low frequencies, and, at best, the likelihood bounds on recombination from linkage disequilibrium are comparable to those found from linkage analysis.

The picture is confused for HD. Although all published studies show markers near D4S95 in disequilibrium with the disease, the most frequent alleles have similar frequencies in the normal and disease samples. The Dutch data may be more informative, because of greater homogeneity of that population, and suggest that the recombination fraction between D4S95 and the HD gene is $<.014$. The real problem is that there are multiple ancestral haplotypes (MacDonald et al. 1992), and no single haplotype has a high current frequency. It is possible that larger sample sizes will help, but clinical variation in the disease (e.g., variable age at onset, severity, and rate of progress) suggests that there are multiple disease mutations. It is also possible that expansions of the trinucleotide repeat associated with the disease occur only on particular haplotypes, and that this is why linkage disequilibrium is ob-

served for D4S95/*Acc*I and D4S95/*Mbo*I. Of course, this phenomenon would invalidate the present model.

The fourth example discussed was diastrophic dysplasia in Finland. The linkage disequilibrium data there are very compelling, and there is no question that markers close to the disease locus have been identified. Hästbacka et al. (1992) gave a moment estimate of cG, and then appealed to the Luria-Delbrück theory to place an upper bound on cG values. Our simulations showed that, by failing to take into account the stochasticity of the growth process, these authors gave an estimate that is too low ~70% of the time and a bound that is less than the parameter value ~40% of the time. We recommend caution when using this bound. In contrast, the likelihood bound is almost always greater than the true value but tends to be conservative.

The likelihood method was easily extended to two marker loci and may help to determine the relative ordering of disease and marker loci. We were able to identify the correct order for markers MET and D7S8 and the CF locus. The analysis depends on the estimated recombination fraction c_{12} between the markers, and could proceed by estimating likelihoods conditional on c_{12}, weighting the results by the likelihoods of c_{12} obtained from linkage studies, and then maximizing the overall likelihood. This approach, which would be superior to the simpler one we used in this paper, would require detailed linkage analyses for marker loci.

Linkage disequilibrium data do not require the identification of many members from disease families, but the conclusions from such data can be ambiguous. There is the danger that diseases such as CF, where linkage disequilibrium can help locate the disease gene, will prove to be the exception, and diseases such as FA will prove to be the norm. We hope that the discussion in this paper will lead to appropriate caution in interpreting linkage disequilibrium data but will lead to useful conclusions when these are warranted.

Copies of a program to carry out the likelihood calculations described in this paper are available from B.S.W. Hästbacka et al. (1994) cloned the DTD gene and found it in the predicted location, 70 kb proximal to the CSF1R gene.

Acknowledgments

This work was supported in part by NIH grant GM45344. Useful comments on a draft of the paper were made by John Drake, Beth Gladen, and anonymous reviewers. Eden Martin verified the numerical results and prepared the program available for distribution.

References

Adam S, Theilmann J, Buetow K, Hedrick A, Collins C, Weber B, Huggins M, et al (1991) Linkage disequilibrium and modification of risk for Huntington disease. Am J Hum Genet 48: 595–603

Andrew SE, Goldberg YP, Kremer B, Squitieri F, Theilmann J, Zeisler J, Telenius H, et al (1994) Huntington disease without CAG expansion: phenocopies or errors in assignment? Am J Hum Genet 54:852–863

Andrew S, Theilmann J, Hedrick A, Mah D, Weber B, Hayden MR (1992) Nonrandom association between Huntington disease and two loci separated by about 3 Mb on 4p16.3. Genomics 13:301–311

Beaudet A, Bowcock A, Buchwald M, Cavalli-Sforza L, Farrall M, King M-C, Klinger K, et al (1986) Linkage of cystic fibrosis to two tightly linked DNA markers: joint report from a collaborative study. Am J Hum Genet 39:681–693

Beaudet AL, Feldman GL, Fernbach SD, Buffone GJ, O'Brien WE (1989) Linkage disequilibrium, cystic fibrosis, and genetic counseling. Am J Hum Genet 44:319–326

Chakravarti A, Buetow KH, Antonarakis SE, Waber PG, Boehm CD, Kazazian HH (1984) Nonuniform recombination within the human β-globin gene cluster. Am J Hum Genet 36:1239–1258

Chamberlain S, Farrall M, Shaw J, Wilkes D, Carvajal J, Hillerman R, Doudney K, et al (1993) Genetic recombination events which position the Friedreich ataxia locus proximal to the D9S15/D9S5 linkage group on chromosome 9q. Am J Hum Genet 52:99–109

Chamberlain S, Shaw J, Rowland A, Wallis J, South S, Nakamura Y, von Gabain A, et al (1988) Mapping of mutation causing Friedreich's ataxia to human chromosome 9. Nature 334:248–250

Cox TK, Kerem B, Rommens J, Iannuzzi MC, Drumm M, Collins FS, Dean M, et al (1989) Mapping of the cystic fibrosis gene using putative ancestral recombinants. Am J Hum Genet 45: A136

Cutting GR, Antonarakis SE, Buetow KH, Kadch LM, Rosenstein BJ, Kazazian HH (1989) Analysis of DNA polymorphism haplotypes linked to the cystic fibrosis locus in North American black and Caucasian families supports the existence of multiple mutations of the cystic fibrosis gene. Am J Hum Genet 14:307–318

de la Chapelle (1993) Disease mapping in isolated human populations: the example of Finland. J Med Genet 30:857–865

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

Estivill X, Farrall M, Williamson R, Ferrari M, Seia M, Giunta AM, Novelli G, et al (1988) Linkage disequilibrium between cystic fibrosis and linked DNA polymorphisms in Italian families: a collaborative study. Am J Hum Genet 43:23–28

Ewens WJ (1979) Mathematical population genetics. Springer, New York

Farrall M, Wainwright BJ, Feldman GL, Beaudet A, Sretenovic Z, Halley D, Simon M, et al (1988) Recombination between IRP and cystic fibrosis. Am J Hum Genet 43:471–475

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

Gilliam TC, Tanzi RE, Haines JL, Bonner TI, Faryniarz AG,

Hobbs WJ, MacDonald ME, et al (1987) Localization of the Huntington's disease gene to a small segment of chromosome 4 flanked by D4S10 and the telomere. Cell 50:565–571

Hanuaer A, Chery M, Fujita R, Driesel AJ, Gilgenkrantz S, Mandel JL (1990) The Freidreich 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

Hästbacka J, de la Chappelle A, Kaitila I, Sistonen P, Weaver A, 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, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, et al (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1073–1087

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

——— (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. Am J Hum Genet 54:705–714

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

Kaplan N, Lewis PO, Weir BS. Age of the ΔF_{508} cystic fibrosis mutation. Nature Genet (in press)

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

———. The use of linkage disequilibrium for estimating the recombination fraction between a marker and a disease gene. In: Donnelly P, Tavaré S (eds) Mathematical population genetics. Springer, New York (in press)

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

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's disease region. Am J Hum Genet 49:723–734

MacDonald ME, Novelletto A, Lin C, Tagle D, Barnes G, Bates G, Taylor S, et al (1992) The Huntington's disease candidate region exhibits many different haplotypes. Nature Genet 1:99–103

Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Giménez J, Reis A, et al (1994) The origin of the major cystic fibrosis mutation (ΔF508) in European populations. Nature Genet 7: 169–175

Pandolfo M, Sirugo G, Antonelli A, Weitnauer L, Ferretti L, Leone M, Dones I, et al (1990) Friedreich ataxia in Italian families: genetic homogeneity and linkage disequilibrium with the marker loci D9S5 and D9S15. Am J Hum Genet 47:228–235

Rodius F, Duclos F, Wrogemann K, Le Paslier D, Ougen P, Billault A, Belal S, et al (1994) Recombinations in individuals homozygous by descent localize the Friedreich ataxia locus inb a cloned 450-kb interval. Am J Hum Genet 54:1050–1059

Romeo G, Menozzi P, Ferlini A, Fadda S, Di Donato S, Uziel G, Lucci B, et al (1983) Incidence of Friedreich ataxia in Italy estimated from consanguineous marriages. Am J Hum Genet 35:523–529

Serre JL, Simon-Bouy B, Mornet E, Jaume-Roig B, Balassopoulou A, Schwartz M, Taillandier A (1990) Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in population genetics. Hum Genet 84:449–454

Sirugo G, Keats B, Fujita R, Duclos F, Purohit K, Koenig M, Mandel JL (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

Skraastad MI, Van de Vosse E, Belfroid R, Höld K, Vegter-van der Vlis M, Sandkuijl LA, Bakker E, et al (1992) Significant linkage disequilibrium between the Huntington disease gene and the loci D4S10 and D4S95 in the Dutch population. Am J Hum Genet 51:730–735

Snell RG, Lazarou LP, Youngman S, Quarrell OW, 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

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. Liss, New York, pp 81–86

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

Whaley WL, Michiles F, MacDonald ME, Romano D, Zimmer M, Smith B, Leavitt J, et al (1988) Mapping of D4S98/S114/S113 confines the Huntington's defect to a reduced region at the telomere of chromosome 4. Nucleic Acids Res 16:11769–11780