Am. J. Hum. Genet. 57:1486–1498, 1995

Are Moment Bounds on the Recombination Fraction between a Marker and a Disease Locus Too Good to Be True? Allelic Association Mapping Revisited for Simple Genetic Diseases in the Finnish Population

N. L. Kaplan[1] and B. S. Weir[2]

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

Summary

In the past several years, allelic association has helped map a number of rare genetic diseases in the human genome. A commonly used upper bound on the recombination fraction between the disease gene and an associated marker is known to be biased downward, so there is the possibility that an investigator could be misled. This upper bound is based on a moment equation that can be derived within the context of a Poisson branching process, so its performance can be compared with a recently proposed likelihood bound. We show that the confidence level of the moment upper bound is much lower than expected, while the confidence level of the likelihood bound is in line with expectation. The effects of mutation at either the marker or disease locus on the upper bounds are also investigated. Results indicate that mutation is not an important force for typical mutation rates, unless the recombination fraction between the marker and disease locus is very small or the disease allele is very rare in the general population. Finally, the impact of sample size on the likelihood bound is investigated. The results are illustrated with data on 10 simple genetic diseases in the Finnish population.

Introduction

Linkage analysis is a powerful method for localizing disease genes to a small region of the genome. To date, upward of 400 genes in the human genome have been mapped using this approach, and ~40 have been successfully cloned (Lander and Schork 1994). The method uses family data, so typically the limited number of available families are not informative for markers close to the disease gene, say within a centimorgan. Therefore, linkage analysis is not particularly useful for fine mapping a gene.

Allelic association mapping (also known as linkage disequilibrium mapping), which relies on population data, is one approach that may help to further localize genes responsible for simple diseases that are relatively rare in the population. These diseases, such as cystic fibrosis (CF), Huntington disease, and diastrophic dysplasia (DTD), are due to a mutation in one gene whose inheritance pattern follows the laws of Mendelian genetics. In contrast, for the more common complex diseases such as diabetes, the disease does not cosegregate with a single locus. The underlying idea of allelic association mapping is that disease chromosomes that descend from the same ancestral mutation should share a common haplotype in a neighborhood of the disease locus, reflecting the haplotype on the ancestral chromosome on which the mutation occurred. Markers near the disease gene are therefore more likely to have an allele in higher frequency in the disease population than in the general population. Hence markers that have an allele in high frequency in a sample of disease chromosomes but in lower frequency in a sample of normal chromosomes are good candidates for being near the disease gene. It was noted by Hästbacka et al. (1992) that the best chance of success with allelic association mapping is when most of the disease chromosomes in the population descend from a single ancestral mutation and the mutation is old enough to allow recombination to break up the ancestral haplotype, but not so old that the maintained neighborhood around the disease locus is too small to be easily detected. These conditions often hold for isolated founder populations that are not very old. A well-studied example is Finland (Nevanlinna et al. 1972; de la Chapelle 1993).

Standard contingency table analyses can be used to compare the marker allele frequency distributions in normal and disease samples. These analyses identify markers that may be close to the disease gene, but do

Received July 11, 1995; accepted for publication September 12, 1995.

Address for correspondence and reprints: Dr. Norman L. Kaplan, National Institutes of Health, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709.

0002-9297/95/5706-0027$02.00

not provide any estimate of how close. Such estimates are necessary when one is trying to clone the gene. Recently, Hästbacka et al. (1992) proposed an estimate of the recombination fraction between a marker and a disease locus that was derived from a simple formula for the expected proportion of the disease chromosomes in the population that carry the ancestral marker allele. They recognized that the distribution of the proportion is skewed (that is, the median is greater than the mean) because of early recombination events, and to compensate for the skewness they used Luria-Delbrück–type arguments to modify the estimate and to obtain what they believed to be approximate bounds.

In a subsequent paper, Kaplan et al. (1995) argued that, if allelic association is detected for a rare human disease, then the disease population is not very old, and they proposed that its growth could be modeled with a Poisson branching process. Adopting a Monte Carlo approach, they simulated disease populations to account for evolutionary variability and estimated confidence bounds from the likelihood of the recombination fraction. Since the moment equation used by Hästbacka et al. can be derived within the context of the model used by Kaplan et al., the confidence levels of the moment bound and the likelihood bound can both be investigated using simulation methods. In particular, Kaplan et al. gave some results indicating that the moment bound tends to underestimate the recombination fraction, suggesting that the Luria-Delbrück approximation does not adequately account for the skewness of the distribution.

Lehesjoki et al. (1993) modified the moment estimate of the recombination fraction to allow for mutation at the disease locus. They were concerned about this issue because, for the genetic disease they were studying, progressive myoclonus epilepsy of the Unverricht-Lundborg type (EPM1), the frequency of the inferred ancestral marker allele in the normal sample was not small. Since their estimate is also based on a moment formula, it too may be biased downward. They presented only the point estimate and alluded to the Luria-Delbrück approximations for confidence bounds. To investigate the effect of recurrent mutation at the disease locus by using the likelihood method, the Poisson branching model needs to be modified. Such a modification is not difficult if it is assumed that the Finnish population is growing exponentially.

To investigate the impact of mutation at the marker locus as well as at the disease locus, requires some additional modifications to the Poisson branching process. Since most of the markers now used are microsatellite markers, with fairly high mutation rates (Weissenbach et al. 1992), it is possible that mutation at the marker might also bias the results. Fortunately, the error is conservative, since ignoring mutation when it is present leads to bounds that are larger than necessary (Lehesjoki et al. 1993). Even so, it is of interest to quantify the bias. Correctly modifying the branching process can be a problem for a microsatellite marker because it requires modeling the process of changes from one marker allele to another, and theories about this process are still speculative. To get around this issue, alleles are collapsed into two categories, ancestral and nonancestral, and it is assumed that mutations from the nonancestral allele to the ancestral allele are so infrequent that they can be ignored. Hence, only mutation away from the ancestral allele is assumed to be relevant.

Sample size considerations can also be investigated with a simulation approach. For most studies the number of disease chromosomes is small, rarely >100, and sometimes as small as 30 (Kestilä et al. 1994). In designing an experiment it would be of value to know how much additional information one could obtain with larger samples. The same is true with the normal sample, especially since the allele frequencies in the normal sample are treated as constant by both Hästbacka et al. (1992) and Kaplan et al. (1995). In some cases the normal sample is as small as 32 (Kestilä et al. 1994), so ignoring sampling error could lead to a bias.

The purpose of this paper is to use the simulation-based likelihood approach proposed by Kaplan et al. (1995) to compare the behavior of the moment bound and the likelihood bound of the recombination fraction, both in the presence and absence of mutation at either the disease or marker locus. The benefits of increasing either the disease or normal sample size are also investigated. To simplify the presentation and minimize the number of simulation parameters, we focus on models suitable for the Finnish population. Also, rather than simulating data, we will demonstrate the behavior of the upper bounds for 10 recently published data sets for simple genetic diseases in Finland.

Methods

To make the paper self-contained, we briefly review the likelihood method proposed by Kaplan et al. (1995). We assume that the disease is relatively rare and is caused by a mutation in a single gene. Suppose a mutation at the disease locus was introduced into the population G generations ago on a chromosome carrying allele M_1 at the marker locus M, and that most if not all the disease chromosomes are descendants of this mutation. In the founder population it is assumed that only one individual carried the mutation. In practice, M_1 is identified as the high-frequency allele in the disease sample. All the other possible alleles are collapsed into one category denoted by M_2. The frequen-

cies of marker allele i in the normal and disease samples are denoted by f_{in} and f_{id} respectively, and their population frequencies by p_{in} and p_{id}, $i = 1,2$. The values of p_{in} are assumed to be constant since the disease mutation is assumed to be of recent origin. Since we will be simulating values of f_{1d}, we denote the observed values of f_{1d} and f_{1n} by $\hat{f}_{1d}$ and $\hat{f}_{1n}$.

No Mutation

All of the diseases we consider are recessive, although the analysis could also apply to late-onset dominant disease such as Huntington disease. Since the disease is in low frequency in the general population, it is reasonable to assume that individuals are either heterozygous at the disease locus (carrier) or homozygous for the normal allele (noncarrier). In Kaplan et al. (1995) it is argued that the disease population cannot be very old, so that modeling the initial growth of the disease population is critical. If all carriers are selectively equivalent, then for a broad class of population genetic models (Ewens 1979), the initial growth of the disease population can be modeled as a Poisson branching process. More specifically, for $i = 1,2$, let

$X_i(t)$ = number of disease chromosomes carrying allele M_i in the tth generation after the introduction of the mutation into the population .

Thus $X_1(0) = 1$, $X_2(0) = 0$, and $X_T(t) = X_1(t) + X_2(t)$ is the number of disease chromosomes in generation t. $X_T(G)$ is the number of disease chromosomes currently in the population. The key assumption is that conditional on $\{X_1(t),X_2(t)\}$, $X_1(t + 1)$ and $X_2(t + 1)$ have independent Poisson distributions with means $m_1(X_1(t), X_2(t))$ and $m_2(X_1(t), X_2(t))$. The functions m_1 and m_2 reflect the assumptions made about recombination and mutation at the two loci. If mutation is ignored at both loci, then for $i = 1,2$

$$m_i[X_1(t), X_2(t)] = (1 + \lambda)[(1 - c)X_i(t) + cX_T(t)p_{in}] , \quad (1)$$

where c is the recombination fraction between the marker and disease loci and λ is a small positive quantity reflecting the growth rate of the entire population as well as the possible selective advantage of carriers over noncarriers. Since the disease mutation has entered the population, it is assumed that $\lambda > 0$.

It is not difficult to show that conditional on $X_T(t)$, $X_1(t)$ has a binomial distribution with parameters $X_T(t)$ and

$$\frac{m_1[X_1(t-1), X_2(t-1)]}{m_1[X_1(t-1), X_2(t-1)] + m_2[X_1(t-1), X_2(t-1)]} .$$

It follows that

$$E\left[\frac{X_1(t)}{X_T(t)} \,\middle|\, X_T(t)\right] = \frac{(1 - c)X_1(t - 1)}{X_T(t - 1)} + cp_{1n} .$$

Taking expectations and iterating leads to the equation

$$E\left[\frac{X_1(G)}{X_T(G)}\right] = (1 - c)^G + [1 - (1 - c)^G]p_{1n} . \quad (2)$$

If we estimate the expectation on the left with $\hat{f}_{1d}$ and p_{1n} with $\hat{f}_{1n}$, then a moment estimate of c is

$$\hat{c} = -\frac{1}{G}\ln\left(\frac{\hat{f}_{1d} - \hat{f}_{1n}}{1 - \hat{f}_{1n}}\right) . \quad (3)$$

This estimate of c is essentially the same one used by Hästbacka et al. (1992) since $\hat{f}_{1n}$ was negligible for DTD. Following Lehesjoki et al. (1993), we write

$$\hat{p}_{\text{excess}} = \frac{\hat{f}_{1d} - \hat{f}_{1n}}{1 - \hat{f}_{1n}} .$$

The statistic p_{excess} was introduced by Bengtsson and Thomson (1981) and recently used by Risch et al. (1995).

The difficulty with using equation (3) to estimate c is that the expectation in equation (2) takes into account the variation resulting from the evolutionary forces responsible for determining the distribution of p_{1d}. To estimate this expectation properly we would need observations from replicate populations, which is not possible. In essence, we are estimating the mean of a random variable with a single observation, and this will work as long as the variance is not large. Since we assume that $X_1(0) = 1$, the distribution of p_{1d} is skewed and the variance is not small. Because of the skewness, a random value of p_{1d} will tend to be larger than its mean, and we would expect equation (3) to lead to underestimates of c. In an attempt to account for the skewness of the distribution of p_{1d}, Hästbacka et al. (1992) proposed the following bound for c, which is based on Luria-Delbrück arguments. The upper moment bound on c, c_H, is the solution of the equation

$$\hat{p}_{\text{excess}} = \frac{c_H}{\lambda}\ln\left[\frac{c_H X_T(G)}{\lambda}\right] - \frac{2c_H}{\lambda} ,$$

where $X_T(G)$ is the current number of disease chromo-

somes in the population. A lower moment bound can be defined analogously, but we are not concerned with its behavior in the present article.

Effects of Mutation

The amount of mutation at the disease locus each generation depends upon the mutation rate as well as the size of the normal population, and to allow for mutation at the disease locus, assumptions need to be made about the growth of the normal population. For isolated founder populations, exponential growth seems like a reasonable assumption, but other models of growth such as a logistic model could easily be implemented. Usually there is an estimate of the current size of the population, for example, for Finland the population is about $N = 5{,}000{,}000$ or $2N = 10{,}000{,}000$ chromosomes (Hästbacka et al. 1992; de la Chapelle 1993). It follows that the expected number of new mutants at the disease locus in the kth generation after the initial settlement of the population is approximately $\mu_D 2N(1+\lambda_o)^{-(G-k)}$, where μ_D is the mutation rate at the disease locus and λ_o is the growth rate of the normal population. For simplicity we assume that $\lambda_o = \lambda$. The m_i now become

$$\begin{aligned} m_i[X_1(t), X_2(t)] &= (1+\lambda)\{(1-c)X_i(t) \\ &\quad + [cX_T(t) + \mu_D 2N(1+\lambda)^{-(G-t)}]p_{in}\} . \end{aligned} \tag{4}$$

The same argument used to justify equation (2) can be used to show that

$$E\left[\frac{X_1(G)}{X_T(G)}\right] \approx \left(1-c-\frac{\mu_D}{q}\right)^G + \left[1-\left(1-c-\frac{\mu_D}{q}\right)^G\right]p_{1n} , \tag{5}$$

where q is the frequency of the disease allele in the overall population. To derive equation (5) it is assumed that

$$\frac{X_T(t)(1+\lambda)^{G-t}}{2N}$$

is constant in time and equals q. Equation (5) is essentially the same result in Lehesjoki et al. (1993). The right-hand side in equation (3) is now an estimate of $c + \mu_D/q$, which shows in particular that ignoring mutation at the disease locus is conservative.

Most markers now typed are microsatellite markers. The large number of alleles is advantageous for linkage studies, but is of less value for studying allelic association since the underlying mutation process for the microsatellite is not well characterized. To simplify the analysis we assume that mutations at the marker locus that change a nonancestral allele to the ancestral one are sufficiently rare that they can be ignored. This allows us to collapse the marker alleles into two categories, ancestral and other, and consider only mutation away from the ancestral allele. With this assumption, the most general forms for the m_i become

$$\begin{aligned} m_1[X_1(t), X_2(t)] &= (1+\lambda)\{(1-c-\mu_M)X_1(t) \\ &\quad + [cX_T(t) + \mu_D 2N(1+\lambda)^{-(G-t)}]p_{1n}\} , \end{aligned}$$

and (6)

$$\begin{aligned} m_2[X_1(t), X_2(t)] &= (1+\lambda)\{(1-c)X_2(t) \\ &\quad + [cX_T(t) + \mu_D 2N(1+\lambda)^{-(G-t)}]p_{2n} + \mu_M X_1(t)\} , \end{aligned}$$

where μ_M is the mutation rate at the marker.

The generalization of equation (5) is straightforward. Indeed,

$$\begin{aligned} E\left[\frac{X_1(G)}{X_T(G)}\right] &\approx \left(1-c-\frac{\mu_D}{q}-\mu_M\right)^G \\ &\quad + \left[1-\left(1-c-\frac{\mu_D}{q}-\mu_M\right)^G\right] \\ &\quad \times \frac{p_{1n}\left(c+\frac{\mu_D}{q}\right)}{\left(c+\frac{\mu_D}{q}+\mu_M\right)} . \end{aligned} \tag{7}$$

Equation (7) does not lead to a simple moment estimate of any linear combination of the parameters unless $c + \mu_D/q \gg \mu_M$. However, it is not hard to show that if we ignore μ_M, then we overestimate $c + \mu_D/q$, which is what we would expect.

Equations (2), (5), and (7) can lead to underestimates of c, since the distribution of p_{1d} is skewed, and so it is possible that inferences about the parameters may be misleading. The simulation approach does not have these problems, so it is of interest to examine the effects of mutation at the marker and disease loci with this approach to see whether the results support the conclusions about the mutation processes from equation (7).

The basic idea of Kaplan et al. (1995) is to estimate the distribution of p_{1d} from simulations and use it to make inferences about c for a given data set. The details of the simulation are given in Kaplan et al. (1995). Con-

ditional on p_{1d}, the number of disease chromosomes in a sample of size k_d carrying the M_1 allele has a binomial distribution with parameters k_d and p_{1d}. Hence, to calculate the likelihood of c, the distribution of p_{1d} is estimated and the binomial probabilities are averaged with respect to this distribution. To remove numerical problems the likelihood is scaled by the binomial probability calculated assuming $p_{1d} = \hat{f}_{1d}$. Having estimated the likelihood, standard likelihood methods involve dropping down 2 from the maximum of the ln(likelihood) to give a support interval for c corresponding to a 95% confidence interval. The upper bound on c determined in this way is denoted by c_M. One can also determine a lower likelihood bound for c. We will not consider its behavior except to say that for many of the markers it is zero and is therefore not very informative.

To carry out simulations, values need to be assigned to the model parameters G and λ. This issue is discussed in detail by Kaplan et. al. (1995), and a method is proposed that exploits prior knowledge about $X_T(G)$, the current size of the disease population. One advantage of restricting our attention to Finland, is that estimates of G and $X_T(G)$ are available from the literature. Indeed, for Finland it is estimated that $G = 100$ generations and $X_T(G) = 10^7$ (Hästbacka et al. 1992; de la Chapelle 1993). Assuming exponential growth, one finds that λ is $\sim$.1 (Hästbacka et al. 1992; Kaplan et al. 1995).

A benefit of using a simulation approach is that the distribution of p_{1d} can be estimated conditional on specified behavior of the underlying disease population. For example, Kaplan et al. (1995) required that the size of the simulated disease population approximate its estimated value. If no condition is placed on the allele frequencies in the normal and disease samples, then for a small fraction of the simulated samples the ancestral allele frequency in the disease sample would be less than its frequency in the normal sample ($f_{1d} < f_{1n}$). These simulations do not affect the likelihood calculations very much, since the associated binomial probability is very small. However, for these samples p_{excess} is negative, and consequently we cannot calculate the moment bound. To avoid this problem, we constrain the simulations in the following additional ways. First, the marker must be polymorphic in the disease sample since all the published data sets have this property. Second, the frequencies of the ancestral allele in the simulated disease population and the disease sample must be larger than $\hat{f}_{1n}$, the observed allele frequency in the normal sample. Since we identify the ancestral allele because of the differences between the frequencies in the two samples, this seems like a reasonable requirement. Finally, we require that the marker allele frequencies in the disease sample be statistically different from the marker allele frequencies in the normal sample, i.e., the χ^2 statistic is significant (>3.84). We do this because we would not even be considering the marker if the associated χ^2 statistic was nonsignificant.

With these constraints, 1,000 values of p_{1d} and f_{1d} were simulated in order to estimate their distributions. An approximate maximum likelihood estimate of c was determined since the likelihood was evaluated only on a mesh of values. In most cases the width of the mesh was .001, but for markers very close to the disease gene ($\hat{f}_{1d}$ close to 1), we used .0002. In all cases the likelihood function was unimodal so the maximum can easily be identified.

Since we can simulate the evolutionary distribution of f_{1d} for any specified value of c, an alternative approach for finding an upper bound on c is to consider the tail probabilities of this distribution. More specifically, a value of c is placed in the confidence set if the probability

$$P(c) = P_c(f_{1d} > \hat{f}_{1d})$$

is not too small. Since $P(c)$ decreases as c increases, we can define the bound

$$c_P = \sup\{c : P(c) > .025\},$$

where the supremum is evaluated on the mesh of values of c.

The evolutionary distribution of f_{1d} can also be used to judge the performance of the other two bounds, c_H and c_M. For the bounds to be reasonable we expect the associated tail probabilities, $P(c_M)$ and $P(c_H)$, to be small.

Effect of Sample Size

To examine sample size effects we adopt the following strategy. The normal sample provides only an estimate of p_{1n}, and we want to examine the effect that sampling error has on the bound c_M. To do this we recalculated c_M using upper and lower 95% confidence bounds on p_{1n}. More specifically, the lower bound on c_M, c_{M-}, was calculated assuming that the frequency of the ancestral allele in the normal sample equals the greater of 0 and $\hat{f}_{1n} - \Delta$ where

$$\Delta = 2\sqrt{\frac{\hat{f}_{1n}(1 - \hat{f}_{1n})}{k_n}}.$$

The upper bound, c_{M+} is defined analogously as the lesser of 1 and $\hat{f}_{1n} + \Delta$.

We have already noted that the disease sample gives us information about the current frequency of the ancestral marker allele in the disease population, which is the single realization of the evolutionary process upon which the confidence-bound calculation is based. In-

creasing the disease sample size improves the confidence bound only to the extent that we decrease the binomial sampling variation and consequently improve our estimate of the frequency of the ancestral marker allele in the disease population. Typically, the disease sample consists of all disease chromosomes that are readily available, and increasing the sample size is not a simple matter. We therefore consider only the effect that doubling the sample size has on c_M. To do this, we assume that the frequency of the ancestral allele in the second sample lies in the 95% confidence interval constructed from the first sample. In this way we can determine a plausible range of values of c_M from the combined sample.

Results

We restrict our attention to 10 simple genetic diseases in the Finnish population. As already noted, Finland has many characteristics favorable to allelic association mapping. In particular, the current population descended from a small group that settled the country ~2,000 years ago (~100 generations if 1 generation is 20 years), (Hästbacka et al. 1992; de la Chapelle 1993). For the simulations we therefore use $G = 100$.

The current population size is ~5 million or 10^7 chromosomes. Assuming exponential growth with an initial population size of 1,000, Hästbacka et al. (1992) gave $\lambda = .085$, although this bound should have been .092. Kaplan et al. (1995) used a different argument based on the growth of the disease population, and estimated $\lambda = .1$. They argued that the likelihood analysis is not sensitive to variation in λ as long as $1 + \lambda$ does not change very much. In this paper we will use $\lambda = .1$.

In table 1 we list 10 simple genetic diseases, and for each disease an estimate of the number of disease chromosomes currently in the Finnish population. We assume that the different numbers of disease chromosomes are, in particular, a consequence of the variability of the evolutionary process. Table 2 contains the relevant data for markers showing an association with each disease. The ancestral marker allele (M_1) is assumed to be the most frequent allele in the disease sample, and all the other alleles are collapsed into one category (M_2). In general, we included in table 2 all markers that were judged to be close to the disease gene using a standard χ^2 statistic. The values of the χ^2 statistic are given in table 2. For DTD there were 11 such markers, and we chose 6 that spanned the region of association. For a few of the markers the frequency of the most frequent allele in the disease sample was lower than the frequency of the collapsed category. This caused problems when we calculated the likelihood, and these markers were excluded from table 2. The values of p_{excess} are also given in table 2.

The χ^2 values range from 3.97 (CF/*pJ3.11*) to 210.80 (DTD/*CSF1R/Eco*RI), whereas the values of p_{excess} lie between .35 (CF/*pJ3.11*) and .95 (e.g., Batten disease [CLN3]/D16S298). In general the χ^2 values and the p_{excess} values are closely related, but this is not always the case. For example, the p_{excess} values for EPM1/*PFKL* and EPM1/*D21S25* are .70 and .71, whereas the χ^2 values are 14.02 and 46.60, respectively. The reason for the big difference in χ^2 values is that the ancestral allele at D21S25 is the common allele in the general population and the ancestral allele at *PFKL* is the rare allele in the normal sample. This is presumably a consequence of the stochastic variability of the evolutionary process. The relationship between the two statistics is also sensitive to sample size. For example, markers CLN5/*D21S162*, CLN1/*HY-TM1*, CLN3/*D16S298*, DTD/*CSF1R/Sty*1, and DTD/*CSF1R/Eco*R1 all have values of $p_{excess} > .9$ ("CLN5" stands for late-infantile neuronal ceroid lipofuscinosis; "CLN1" stands for infantile neuronal ceroid lipofuscinosis), while the χ^2 values range from 26.3 (CLN5/*D21S162*) to 210.8 (DTD/*CSF1R/Eco*R1). The difference in these two χ^2 values is caused by the sample size: 25 disease, 25 normal versus 158 disease, 128 normal.

In table 3 the three bounds, c_H, c_M, and c_P, are given for all the markers as well as the associated tail probabilities. The values of c_H are all $<.01$, and in some cases are considerably smaller, e.g., for CLN1/*HY-TM1*, $c_H = .0011$, and DTD/*CSF1R/Sty*1, $c_H = .0012$. If we use the usual conversion of 1 cM = 1,000 kb, then c_H indicates that these two markers are within 100 kb of the disease locus. The associated tail probabilities for these markers, .25 and .30, however, are not small. In fact, all of the tail probabilities in table 3 for c_H are large, suggesting that the c_H bound does not provide a meaningful upper bound. The smallest tail probability is .18 (CLN5/*D13S162*), while the largest is .64 (DTD/*PDGFRB/Bg1*1). Hence, it appears that the Luria-Delbrück correction proposed by Hästbacka et al. (1992) is not adequate. For two of the markers, DTD/*RPS14* and CF/*pJ3.11*, the moment estimate of c was actually greater than the bound, indicating that the algorithm for calculating c_H is not appropriate if p_{excess} is too small ($p_{excess} = .35$ and .34). Also included in table 3 are the maximum likelihood estimates of c (c_{ML}). They are close to the corresponding values of c_H, suggesting that it may be more appropriate to think of c_H as a point estimate of c rather than as an upper bound of c.

The likelihood bound, c_M, for each of the markers is larger than the corresponding value of c_H, and for many markers, c_M is about twice the value of c_H. We are not sure why this is so, but this does give a quick and reasonable estimate of c_M. The associated tail probabilities for c_M are consistently below the nominal .025 value.

Table 1

Simple Genetic Disease in the Finnish Population for Which There Are Markers Showing Allelic Association

Disease	Estimated No. of Disease Chromosomes in the Finnish Population × 10^{-4}	Reference
DTD	6.0	Hästbacka et al. 1994
EPM1	7.1	Lehesjoki et al. 1993
Cartilage-hair hypoplasia (CHH)	6.6	Sulisalo et al. 1994
CLN3	6.9	Mitchison et al. 1995
Congenital nephrotic syndrome (CNF)	11.0	Kestila et al. 1994
CLN1	7.1	Hellsten et al. 1993
APECED	6.3	Aaltonen et al. 1994
CLN5	2.1	Savukoski et al. 1994
SD	4.2	Hattaja et al. 1994
CF	6.1	Ramsay et al. 1993

However, the values of c_P indicate that the c_M bounds are not excessively conservative. In most cases c_P is within .001 of c_M.

The disease genes for CF and DTD have been identified, so it is of interest to see how the bounds perform in these cases. The DTD gene is ~70 kb from *CSF1R* (Hästbacka et al. 1994), whereas for *CSF1R/Sty*1 c_H = .0012 and c_M = .0026. In this case the likelihood bound identifies a target region that is ~250 kb proximal to *CSF1R*, while the moment bound indicates a region about half that size. However, it is inappropriate to use this one example to justify the use of the moment bound.

The CF markers, *pXV-2C* and *pKM.19*, are ~250 kb and ~200 kb, respectively, from the CF gene (Kerem et al. 1989), and so the associated recombination fractions are ~.0025 and ~.002. The c_H bounds are .0028 and .0051, whereas the c_M bounds are .009 and .014. In both cases the bounds incorrectly imply that *pKM.19* may be closer to the gene than *pXV-2C*. The associated χ^2 values and p_{excess} values also support the incorrect order. In contrast, CF data for these markers for many of the European countries are consistent with the correct order of the markers (Kaplan et al. 1995). The Finnish data are different because the frequency of the ancestral *pKM.19* allele is lower in the disease sample and higher in the normal sample than in the other European countries. One explanation for this difference is the small sample size (38 disease and 37 normal). It is possible that with larger samples, the bounds would be consistent with the correct order. Then again, it is also possible that the evolutionary history of the Finnish population is different from that of the rest of Europe.

The ancestral allele frequencies for *pXV-2C* are consistent with samples from other parts of Europe (Kaplan et al. 1995). The large value of c_M, .009, for *pXV-2C* in table 3 reflects the young age of the Finnish population since the critical parameter is the product cG. Since Kaplan et. al. (1995) used G = 200 in their analysis, we need to halve .009 to make the proper comparison. When we do this we find that .0045 is in line with the values in table 2 in Kaplan et al. Similarly, if we were to halve c_H, then we obtain .0014 which is substantially below the actual value .0025 (Kerem et al. 1989).

To explore the effects of mutation at the marker and disease loci, we calculated c_M for different values of μ_M and μ_D. The values of μ_D considered were 10^{-6} and 10^{-5} (Sulisalo et al. 1994; Mitchison et al. 1995). Two values of μ_M, 10^{-4}, and 10^{-3}, were examined for microsatellite markers (Weissenbach et al. 1992). In table 4, values of c_M are given for one marker for each of the 10 diseases. The first nine markers are microsatellites, but the last is an RFLP and so only mutation at the disease locus was considered. The results are consistent with the moment-based predictions. In particular, if the value of c_M calculated assuming $\mu_M = \mu_D = 0$ is much larger than $\mu_M + \mu_D/q$, then the effect of mutation on c_M is negligible. Alternatively, if c_M is smaller than $\mu_M + \mu_D/q$, then allowing for mutation will lead to a reduction in c_M. For example, for markers 5 and 7, the values of c_M change from .013 and .014 to .011, while for markers 4 and 6, the values of c_M change from .0030 and .0024 to .0016 and .0008, respectively.

This analysis assumes that one cannot identify disease chromosomes carrying nonancestral marker alleles that have arisen because of mutation at either the marker or disease locus rather than recombination. If these chromosomes can be identified, then it is preferable to omit

Table 2

Sample Data for Markers Associated with the Disease Gene

Disease and Marker	Disease M_1	Disease M_2	Normal M_1	Normal M_2	p_{excess}	χ^2
DTD:						
D5S372	93	61	16	103	.54	61.68
*CSF1R/Sty*1	151	7	34	93	.94	146.31
CSF1R/TAGA[a]	144	6 (14)	46	82	.94 (.86)	115.16 (96.63)
*CSF1R/Eco*RI	150	8	12	116	.93	210.80
*PDGFRB/Bgl*I	94	47	36	87	.53	37.76
RPS14	99	51	57	63	.35	9.35
EPM1:						
D21S141	54	20	11	41	.66	32.84
PFKL	56	20	7	47	.70	46.60
*PFKL/Kpn*I	65	11	33	22	.64	11.03
D21S25	67	9	31	21	.71	14.02
D21S171	50	26	11	41	.57	26.89
CHH:						
D9S163	109	19	41	57	.75	46.67
D9S50	69	56	10	88	.50	48.62
CLN3:						
D16S288	33	11	3	41	.73	42.31
D16S299	47	7	13	41	.83	43.35
D16S298	52	2	16	38	.95	51.46
SPN	36	8	16	28	.71	18.80
CNF:						
D19S224	21	11	4	28	.61	18.97
D19S220	22	10	2	30	.66	26.67
CLN1:						
HY-TM1	76	4	2	78	.95	136.99
L-MYC	72	6	39	35	.84	30.24
D1S62	69	11	39	40	.73	24.82
APECED:						
D21S49	17	11	0	28	.61	24.41
D21S171	18	10	2	26	.62	19.91
CLN5:						
D13S160	23	4	1	26	.85	36.30
D13S162	23	2	5	20	.90	26.30
SD:						
D6S286	38	8	15	31	.74	23.54
CF:						
pXV-2c	35	3	18	19	.84	17.08
pKM.19	31	7	15	22	.69	13.31
pMP6d-9	32	5	13	20	.78	16.86
pG2	36	2	27	10	.80	6.61
pJ3.11	21	17	12	25	.34	3.97

[a] For this polymorphic locus the data omit chromosomes having nonancestral marker alleles that appear to have arisen from mutation rather than recombination. The data before omission are in parentheses.

them from the data and calculate c_M ignoring mutation. For example, Hästbacka et al. (1994) found that 8 of the 14 nonancestral DTD chromosomes at the *CFS1R*/TAGA microsatellite marker retained rare ancestral alleles at flanking markers. If these eight chromosomes are omitted from the data and mutation is ignored ($\mu_D = \mu_M = 0$), then $c_M = .0028$. In an earlier paper Hästbacka et al. (1992) estimated the mutation rate at *CSF1R*/TAGA to be on the order of .0004. The value of c_M for the nonadjusted data with $\mu_M = .0004$ is .004.

We next examined the effect of the size of the normal sample on c_M. In table 5 upper and lower bounds on c_M are given for the 10 markers in table 4. For ease of presentation, the markers are arranged by increasing values of Δ. As expected, larger Δ values give greater effect of sampling error on c_M. However, if c_M is small,

Table 3

The Three Upper Bounds, c_H, c_M, and c_P on the Recombination Fraction between the Marker and Disease Gene, and the Associated Tail Probabilities

Disorder and Marker	c_H	$P(c_H)$	c_{ML}[a]	c_M	$P(c_M)$	c_P
DTD:						
D5S372	.70	.54	.8	1.3	.005	1.20
CSF1R/Sty1[b]	.12	.30	.10	.26	.009	.22
CSF1R/TAGA[b,c]	.13	.25	.12	.28	.013	.24
CSF1R/TAGA	.25	.31	.24	.46	.012	.37
*CSF1R/Eco*RI[b]	.12	.29	.10	.22	.020	.22
*PDGFRB/Bgl*I	.71	.64	.9	1.5	.008	1.40
RPS14	...[d]		1.6	>2		>2
EPM1:						
D21S141	.55	.45	.6	1.1	.007	1.0
PFKL	.49	.38	.5	1.0	.002	.9
*PFKL/Kpn*I	.57	.64	.7	1.8	.021	1.7
D21S25	.48	.49	.6	1.3	.015	1.2
D21S171	.66	.48	.9	1.4	.008	1.3
CHH:						
D9S163	.43	.45	.5	.9	.007	.8
D9S50	.76	.59	1.0	1.5	.005	1.4
CLN3:						
D16S288	.45	.38	.4	.9	.007	.8
D16S299	.30	.36	.3	.7	.002	.6
D16S298[c]	.11	.19	.10	.32	.010	.28
SPN	.47	.38	.5	1.2	.005	1.1
CNF:						
D19S224	.58	.51	.6	1.3	.014	1.2
D19S220	.50	.42	.5	1.1	.005	1.0
CLN1:						
HY-TM1[b]	.11	.25	.10	.24	.013	.20
L-MYC	.29	.37	.3	.7	.012	.6
D1S62	.45	.47	.5	1.1	.010	.9
APECED:						
D21S49	.63	.44	.6	1.3	.014	1.2
D21S171	.61	.47	.7	1.4	.010	1.3
CLN5:						
D13S160	.34	.25	.3	.8	.010	.7
D13S162	.24	.18	.2	.7	.010	.6
SD:						
D6S286	.46	.40	.5	1.1	.010	1.0
CF:						
pXV-2c	.28	.30	.3	.9	.013	.7
pKM.19	.51	.50	.6	1.4	.014	1.3
pMP6d-9	.39	.37	.4	1.1	.013	.9
pG2	.34	.37	.4	1.7	.064	1.7
pJ3.11	...[d]			>2		>2

NOTE.—The entries in the table are for $c_M \times 10^2$.

[a] c_{ML} = maximum-likelihood estimate of c.

[b] The mesh size for these markers was .0002. For all other markers the mesh size was .001.

[c] Disease chromosomes omitted having nonancestral marker alleles that appear to have arisen from mutation rather than recombination.

[d] Ellipses (...) mean that c_H is less than the point estimate.

then even large Δ values have minimal effect, e.g., marker 4. The results in table 5 imply that $k_n = 100$ is a reasonable normal sample size since in this case Δ is always $<.1$.

In table 6 bounds on c_M are given assuming the size of the disease sample is doubled. If c_M for the original sample is large, then increasing the sample size decreases the bound marginally and one is still left with a large

region to explore. On the other hand; if the original c_M is small, e.g., markers 1, 4, and 6, then increasing the sample size does offer some potential decrease in the size of the target region. The effect is not substantive however, and the additional cost of obtaining a larger sample may not be justified.

Discussion

Allelic association data can provide information that is useful for fine mapping a disease gene. Unlike linkage mapping, which relies on recombination events in families, allelic association mapping exploits recombination events that occur in the evolutionary history of the disease. This approach has the best chance of success for relatively rare, simple genetic diseases that are not very old and for which most of the disease chromosomes descend from just a few ancestral mutations. Many of the examples where allelic association has provided useful information are for simple diseases from isolated founder populations such as Finnish (de la Chapelle 1993), Ashkenazi Jew (Motulsky 1995) and Louisiana Acadian (Sirugo et al. 1992). In these populations the disease mutation is usually introduced with the founders, and subsequently increases in frequency. If the population is not very old, then recombination will not have sufficient opportunity to completely break up the ancestral haplotype and the disease chromosomes exhibit a common haplotype in a neighborhood of the disease mutation. Allelic association mapping can also succeed for diseases in nonfounder populations. The most striking example is CF. In this case, ~70% of CF chromosomes worldwide descend from a single three-base deletion, and the haplotype of the ancestral chromosome in a neighborhood of the mutation has remained relatively intact (Kerem et al. 1989).

If the mutation is of recent origin, then its initial growth can be modeled as a Poisson branching process. Within the context of this model three upper confidence bounds on the recombination fraction are discussed: the moment bound c_H, the maximum-likelihood bound c_M and the tail-probability bound c_P. In table 3 these bounds are compared for 10 simple genetic diseases in the Finnish population. The moment bound, c_H, is always $<.01$ (<1 cM), and in many cases is much less. Since the mean of the distribution of p_{1d} is less than the median, it is very possible that c_H is biased downward despite the attempts to adjust for the skewness. The large values of $P(c_H)$ in table 3 support this conclusion. Underestimating the value of c is a serious error to make when devising a mapping strategy and one should be very cautious in the use of c_H.

The likelihood bound c_M is often about twice as large as c_H and may be on the conservative side since most of the values of $P(c_M)$ are slightly less than the nominal value .025. Overestimating c from a mapping perspective is clearly the preferable error. The error does not seem to be excessive, since c_M and c_P typically differ by no more than .001.

The results in table 4 argue strongly for ignoring mutation at either of the two loci, unless the disease is very rare ($q < .001$) or there is some evidence that either of the two mutation rates is unusually large (e.g., $\mu_D > 10^{-5}$ or $\mu_M > 10^{-3}$). The predictions based on moment considerations also hold for c_M (table 4). For exam-

Table 4

The Effect of Mutation at Either the Disease or Marker Locus on the Value of c_M

μ_M and Marker	μ_D .0	10^{-6}	10^{-5}
.0:			
1. DTD/*CSF1*/TAGA[a]	.28	.26	.14
2. EPM1/*PFKL*	1.0	.9	.8
3. CHH/*D9S163*	.9	.9	.7
4. CLN3/*D16S298*	.30	.30	.18
5. CNF/*D19S224*	1.3	1.3	1.2
6. CLN1/*HY-TM1*	.24	.22	.12
7. APECED/*D21S171*	1.4	1.4	1.2
8. CLN5/*D13S160*	.8	.7	.4
9. SD/*D6S286*	1.1	1.1	.9
10. CF/*pKM.19*[b]	1.4	1.4	1.2
10^{-4}:			
1. DTD/*CSF1*/TAGA	.26	.24	.14
2. EPM1/*PFKL*	.9	.9	.8
3. CHH/*D9S163*	.9	.9	.7
4. CLN3/*D16S298*	.28	.30	.18
5. CNF/*D19S224*	1.3	1.3	1.2
6. CLN1/*HY-TM1*	.22	.20	.10
7. APECED/*D21S171*	1.4	1.4	1.2
8. CLN5/*D13S160*	.8	.7	.4
9. SD/*D6S286*	1.1	1.1	.9
10^{-3}:			
1. DTD/*CSF1*/TAGA	.12	.12	.10
2. EPM1/*PFKL*	.8	.8	.7
3. CHH/*D9S163*	.7	.7	.5
4. CLN3/*D16S298*	.16	.18	.16
5. CNF/*D19S224*	1.2	1.2	1.1
6. CLN1/*HY-TM1*	.14	.12	.08
7. APECED/*D21S171*	1.3	1.3	1.1
8. CLN5/*D13S160*	.7	.6	.4
9. SD/*D6S286*	1.0	.9	.7

NOTE.—Included in this table are results for only one marker for each disease. The entries in the table equal $c_M \times 10^2$. The mesh size was .001 except for markers 1, 4, and 6. For these three the mesh size was .0002.

[a] Disease chromosomes omitted having nonancestral marker alleles that appear to have arisen from mutation rather than recombination.

[b] This marker is an RFLP, so mutation at the marker was not considered.

Table 5

Effect of Normal Sample Size on c_M

Marker	Δ	k_n	c_{M-}	c_M	c_{M+}
6. CLN1/*HY-TM1*	.04	80	.24	.24	.26
1. DTD/*CSF1*/TAGA[a]	.08	128	.24	.28	.32
8. CLN5/*D13S160*	.08	27	.8	.8	.9
2. EPM1/*PFKL*	.09	54	.8	1.0	1.1
7. APECED/*D21S171*	.09	28	1.3	1.4	1.6
3. CHH/*D9S163*	.10	98	.7	.9	1.1
4. CLN3/*D16S298*	.12	54	.28	.30	.40
5. CNF/*D19S224*	.12	32	1.1	1.3	1.8
9. SD/*D6S286*	.14	46	.8	1.1	1.6
10. CF/*pKM.19*	.16	38	1.0	1.4	1.6

NOTE.—The number of the marker is the same as in table 4. See text for definition of Δ, c_{M-}, and c_{M+}. Values of c_{M-}, c_M, and c_{M+} are multiplied by 100. The mesh size was .001 except for markers 1, 4, and 6. For these three the mesh size was .0002.

[a] Disease chromosomes omitted having nonancestral marker alleles that appear to have arisen from mutation rather than recombination.

ple, if $\mu_D = 10^{-6}$ and $q > .002$, then $\mu_D/q < .0005$, and so ignoring mutation at the disease locus introduces an error <.001. Even if $\mu_D = 10^{-5}$, the error is at most .005 and could be much less if q is >.002. Ignoring mutation at the marker locus introduces an error that is approximately equal to μ_M. Hence, mutation rates on the order of .001 or smaller can typically be ignored for markers where the estimated value of c_M ignoring mutation is much larger than .001, e.g., EPM1/*PFKL*, APECED/*D21S171*, and CLN5/*D13S160*.

When planning an allelic association study, sample size is an important consideration. The results in table 5 indicate that the normal sample size, k_n, should be large enough to guarantee that $\Delta < .1$. Since $\Delta \leq 2\sqrt{0.25/k_n}$, $k_n = 100$ is a conservative normal sample size. The size of the disease sample is typically not decided a priori, since disease chromosomes are difficult to obtain. Often the investigator has access to a limited number of disease chromosomes and as many of these are typed as possible. The question of interest is whether the additional work and cost to enlarge the sample can be justified. Our answer to this question is based on the behavior of c_M. In particular, we focus on how c_M changes if the sample size is doubled. For the rare diseases considered here, doubling the sample size is probably the most one could expect. Increasing the sample size may alter $\hat{f}_{1d}$, and so in table 6 we calculated c_M for reasonable upper and lower bounds on $\hat{f}_{1d}$ for the combined sample. The results indicate that increasing the sample size has marginal value, but might be informative if c_M for the original sample is small. A prudent strategy would be to type as many disease chromosomes as is practical, and with values of $\hat{f}_{1d}$ and $\hat{f}_{1n}$ in hand, assess whether it is worthwhile trying to enlarge either of the samples.

For many of the diseases in table 1 the disease chromosomes were haplotyped for markers showing a strong association with the disease. These data are important because they provide support for the basic hypothesis that most of the disease chromosomes descended from a single mutation. If there is no dominant haplotype, such as with Huntington disease (MacDonald 1992), then one must be very cautious when using the allelic association data. Fortunately, none of the genetic diseases considered here show multiple haplotypes.

One would hope that haplotype data could be used to help fine map a disease gene. Ramsey et al. (1993) suggested a contingency table procedure that relied on inferring the ancestral disease haplotype so that recombinant disease bearing chromosomes could be identified. Kaplan et al. (1995) proposed a likelihood approach that did not require inferring the ancestral haplotype. Both methods led to the correct conclusion that the CF gene lies telomeric to both markers pXV-2c and pKM-19 (Kerem et al. 1989). Ramsey et al. (1993) also suggested an empirical approach for using multiple marker (>2) haplotype data for fine mapping a disease gene. The statistical methods for analyzing haplotype data are far from definitive, and additional work needs to be done in this important area.

If one knows the relative positions of the markers, then plotting a measure of association, such as p_{excess} or c_M, as a function of the location of the marker can in some cases reveal an obvious gradient and provide information about the location of the disease gene. In these cases one would begin looking for the disease gene

Table 6

Effect of Disease Sample Size on c_M

Marker	$2 \times k_d$	c_{M-}	c_M	c_{M+}
8. CLN5/*D13S160*	54	.5	.8	1.0
7. APECED/*D21S171*	56	.9	1.4	1.5
5. CNF/*D19S224*	64	.9	1.3	1.5
10. CF/*pKM.19*	76	.8	1.4	1.7
9. SD/*D6S286*	92	.7	1.1	1.2
4. CLN3/*D16S298*	108	.18	.30	.40
2. EPM1/*PFKL*	152	.8	1.0	1.1
6. CLN1/*HY-TM1*	160	.14	.24	.28
3. CHH/*D9S163*	256	.7	.9	1.0
1. DTD/*CSF1*/TAGA[a]	316	.18	.28	.30

NOTE.—The number of the marker is the same as in table 4. See text for definition of c_{M-}, and c_{M+}. Values of c_{M-}, c_M, and c_{M+} are multiplied by 100. The mesh size was .001 except for markers 1, 4, and 6. For these three the mesh size was .0002.

[a] Disease chromosomes omitted having nonancestral marker alleles that appear to have arisen from mutation rather than recombination.

in a neighborhood of the marker showing the largest measure of association. Two examples where this approach worked were DTD (Hästbacka et al. 1993) and CF (Kerem et al. 1989). It is informative if there is a gradient for the measure of association, but having a gradient may be due to "evolutionary luck" and be more the exception than the rule.

It is of interest to compare the results of the present article with others in the literature. Several authors give moment estimates of c that are too small because they ignore the frequency of the ancestral marker allele in the normal sample. Hästbacka et al. (1992) did this for markers near DTD because the allele frequencies in the normal sample were so low. However, for Salla disease (SD)/*S286* and CLN5/*D13S162* the frequencies of the ancestral allele in the normal sample are not negligible and cannot be ignored (.33 and .22, respectively). Also, in both cases the sample sizes are quite small and so sampling error could be an issue (for both markers $\Delta = .14$).

The authors studying autoimmune polyglandular disease type 1 (APECED) (Aaltonen et al. 1994) argue that their estimate of c_H should be smaller because of new mutants ($\mu_D > 0$). They suggest that ~7% of the APECED population are due to new mutations. It is not hard to show (Hästbacka et al. 1992) that the fraction of the APECED population due to new mutations is approximately $G\mu_D/q$. Since $\mu_D/q = .07/100 = .0007$ and $c_M = .014$ ignoring mutation, the error introduced in c_M due to $\mu_D > 0$ is negligible.

The strongest claim regarding a bound on c was made by Mitchison et al. (1995) for a marker near CLN3. These authors claim that the recombination fraction between CLN3 and the marker *D16S298* is <.00014, implying that the gene is within 14 kb of the marker. The maximum likelihood bound, c_M, assuming $\mu_D = 5 \times 10^{-6}$ is .0024 which is ~17 times the moment estimate. The value of P(.00014) is ~.28, supporting our contention that one should be cautious about using such a small bound. Only cloning the gene will reveal the true distance.

Finally, it is important to keep in mind that allelic association mapping can be used for relatively rare simple genetic diseases and that generalizing it to complex diseases may not be appropriate. The strength of this approach is that we can exploit the parametric population genetic model to estimate the recombination fraction and move beyond traditional hypothesis testing. For complex diseases the Poisson branching model probably does not hold since susceptibility alleles may not be rare or of recent origin. Hence it is not currently clear how to exploit the evolutionary history of the disease population as was so successfully done for simple genetic diseases.

Acknowledgments

This work was supported in part by NIH grant GM45344. Useful comments on a draft of the paper were made by Beth Gladen, Jack Bishop, and anonymous reviewers.

References

Aaltonen J, Björnses P, Sandkuijl L, Perheentupa J, Peltonen L (1994) An autosomal locus causing autoimmune disease: autoimmune polyglandular disease type 1 assigned to chromosome 21. Nat Genet 8:83–87

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

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

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

Haataja L, Schleutker J, Laine AP, Renlund M, Savontaus ML, Dib C, Weissenbach J, et al (1994) The genetic locus for free sialic acid storage disease maps to the long arm of chromosome 6. Am J Hum Genet 54:1042–1049

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. Nat Genet 2:204–211

Hästbacka J, de la Chapelle A, Mahtani MM, Daly, M Hamilton BA, Kusumi K, Trivedi B, et al (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1073–1087

Hellsten E, Vesa J, Speer MC, Mäkelä TP, Järvelä I, Alitalo K, Ott J, et al (1993) Refined assignment of the infantile neuronal ceroid lipofuscinosis (INCL, CLN1) locus at 1p32): incorporation of linkage disequilibrium in multipoint analysis. Genomics 16: 720–725

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

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

Kestilä M, Männikkö M, Holmberg C, Gyapay G, Weissenbach J, Savolainen ER, Peltonen L, et al (1994) Congenital nephrotic syndrome of the Finnish type maps to the long arm of chromosome 19. Am J Hum Genet 54:757–764

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lehesjoki AE, Koskinieme M, Norio R, Tirrito S, Sistonen P, Lander E, de la Chapelle (1993) Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution. Hum Mol Genet 8:1229–1233

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. Nat Genet 1:99–103

Mitchison H, O'Rawe AM, Taschner PM, Sandkuijl LA, Santavuori P, de Vos N, Breuning MH, et al (1995) Batten disease gene: linkage disequilibrium mapping in the Finnish population, and analysis of European haplotypes. Am J Hum Genet 56:654–662

Motulsky AG (1995) Jewish diseases and origins. Nat Genet 999–101

Nevanlinna HR (1980) Genetic markers in Finland. Haematologica 13:65–74

Ramsey M, Williamson R, Estivill X, Wainwright BJ, Ho MF, Halford S, Kere J, et al (1993) Haplotype analysis to determine the position of a mutation among closely linked markers. Hum Mol Genet 2:1007–1014

Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, Singer B, Fahn S, et al (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small population. Nat Genet 9:152–159

Savukoski M, Kestilä M, Williams R, Järvelä I, Sharp J, Harris J, Santavuori P, et al (1994) Defined chromosomal assignment of CLN5 demonstrates that at least four genetic loci are involved in the pathogenesis of human ceroid lipofuscinoses. Am J Hum Genet 55:695–701

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 micro-satellite-generated extended haplotypes. Am J Hum Genet 50: 559–566

Sulisalo T, Klockars J, Mäkitie O, Francomano, de la Chapelle A, Kaitila I, Sistonen P (1994) High-resolution linkage-disequilibrium mapping of the cartilage-hair hypoplasia gene. Am J Hum Genet 55:937–945

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau, Vaysseix G, et al (1992) A second generation linkage map of the human genome. Nature 359:794–801