LINKAGE ANALYSIS FOR RECESSIVE AND PARTIALLY RECESSIVE TRAITS USING AFFECTED SIBLING PAIRS AND INBRED INDIVIDUALS

by

Eleanor Feingold and David O. Siegmund

TECHNICAL REPORT NO. 4
December 1995

Prepared Under Grant NIH 1-R01-HG00848
For The National Institutes of Health

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Linkage Analysis for Recessive and Partially Recessive Traits Using Affected Sibling Pairs and Inbred Individuals

by

Eleanor Feingold and David O. Siegmund

TECHNICAL REPORT NO. 4
December 1995

Prepared Under Grant NIH 1-R01-HG00848
For The National Institutes of Health

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Linkage Analysis for Recessive and Partially Recessive Traits Using Affected Sibling Pairs and Inbred Individuals

Eleanor Feingold, Department of Biostatistics, Emory University
1518 Clifton Rd. Rm. 314
Atlanta, GA 30322
phone: 404-727-9890, fax: 404-727-1370
e-mail: feingold@sph.emory.edu

David O. Siegmund, Department of Statistics, Stanford University

Running head: Linkage of recessive traits
Key Words: genetic mapping, linkage analysis, sibling pairs, genome scans, inbreeding
Summary

Markov chain and Gaussian process models are developed for whole genome scans to detect linkage of recessive and partially recessive traits using sibling pairs or inbred individuals. For the case of sibling pairs different statistics designed for different degrees of recessiveness are compared, and sibling pairs are compared to more distant relatives, such as cousins. For inbred individuals, the power of offspring of different types of matings is evaluated, and compared to sibling pairs. Over a broad range of trait etiologies, sibling pairs are more powerful than inbred individuals, but for traits caused by very rare alleles, particularly those exhibiting locus heterogeneity and a low frequency of phenocopies, inbred individuals can be much more powerful.
Introduction

The purpose of this paper is to explore strategies for mapping recessive or partially recessive traits using affected sibling pairs and inbred individuals. The context we consider is a whole-genome search for one or more loci, based on identity-by-descent (IBD) methods. Data for such a search would come from a dense set of polymorphic markers, or, in theory, from a method such as Genomic Mismatch Scanning (Nelson et al. 1993) or Representational Difference Analysis (Lisitsyn et al. 1993) that gives whole-genome IBD information in a single procedure. With whole-genome IBD data for sibling pairs and whole-genome homozygosity by descent (HBD) data for inbred individuals, siblings and inbred individuals can be analyzed very similarly, and it is timely to explore overall strategies for using both kinds of data. Feingold (1993), Feingold et al. (1993), Lander and Schork (1994), and Dupuis et al. (1995) described statistical methods for analyzing whole-genome IBD data for affected relative pairs. The essence is to model the number of pairs sharing IBD at any given point on the chromosome as a stochastic process, with the “time” parameter being distance along the chromosome. The methods are applicable to a wide variety of simple and complex modes of inheritance (cf. Dupuis et al. 1995), but for the important case of sibling pairs, most of this literature has concentrated on additive traits (those with zero dominance variance). In this paper, we extend the previous methods to traits with non-zero dominance variance using (1) affected sibling pairs with full information about the number of alleles that are IBD and (2) affected inbred individuals with HBD.
information. The main issue for sibling pairs is choosing the best test statistic
given that the mode of inheritance (i.e., the amount of dominance variance) is likely
to be known only approximately. For inbred individuals, the issue is what types of
matings give the most mapping power. We also compare the utility of inbred
individuals and sibling pairs for various trait etiologies. We begin our discussions
by developing Markov chain and Gaussian models that are necessary to do the
analyses. The Markov chains give very accurate results for small sample sizes.
For larger sample sizes, the Gaussian processes are approximations to the Markov
chains. They are more amenable to complex computations and thus give insight
into questions that can be too difficult to answer in the Markov chain context. All
of the models assume the Haldane mapping function, equal lengths of the male and
female maps, and dense, completely informative markers yielding continuous
information about regions of IBD and HBD. With some cost in complexity, all of
those assumptions can be weakened.

Sibling Pairs

Markov Chain Models

The basic model for sibling pairs was described in Feingold (1993). For
the ith sibling pair, let $X_{t}^{(i)} = (X_{0,t}^{(i)}, X_{1,t}^{(i)}, X_{2,t}^{(i)})$ be the vector-valued Markov chain
giving the IBD status of the pair at point $t$ on the chromosome. The component
$X_{k,t}^{(i)}$ equals one when the pair shares $k$ alleles IBD and equals zero otherwise. If
there are \( N \) independent sibling pairs, the chains can be added together to get

\[
X_t = \sum_{i=1}^{N} X^{(i)}_t = \left( \sum_{i=1}^{N} X^{(i)}_{0}, \sum_{i=1}^{N} X^{(i)}_{1}, \sum_{i=1}^{N} X^{(i)}_{2} \right) = (X_{0,t}, X_{1,t}, X_{2,t}),
\]

which gives the number of pairs that share zero, one, and two alleles IBD at point \( t \). The three components of the vector must, of course, sum to \( N \). The process \( X_t \) is also a Markov chain. If there is no gene on the chromosome linked to the trait, \( X_t \) has for all \( t \) a multinomial distribution with parameters \( (N, 1/4, 1/2, 1/4) \), although the values of \( X_t \) fluctuate along the length of the chromosome. If there is a gene linked to the trait, the probabilities for the multinomial will be different near the point where the gene lies, with more pairs sharing two alleles IBD and fewer sharing zero. If there is only one linked gene on the chromosome, the probabilities at the trait locus can be parameterized as

\[
(1 - \alpha)/4, \quad (1 - \delta)/2, \quad (1 + \alpha + 2\delta)/4
\]

for the probabilities of finding 0, 1, or 2 alleles IBD, respectively. For any other point \( t \) on the chromosome, the above probabilities are respectively

\[
\frac{1}{4} \left[ 1 - (\alpha + \delta) e^{-4|t-r|} + \delta e^{-8|t-r|} \right], \quad \frac{1}{2} \left[ 1 - \delta e^{-8|t-r|} \right], \quad \text{and}
\]

\[
\frac{1}{4} \left[ 1 + (\alpha + \delta) e^{-4|t-r|} + \delta e^{-8|t-r|} \right],
\]

where \( r \) is the point location of the trait locus and \( \lambda \) is the crossover rate per unit
of distance \( t \) (e.g. 0.01/cM). The values of \( \alpha \) and \( \delta \) depend on the specifics of the trait etiology. The parameter \( \alpha \) must lie between zero and one, and the parameter \( \delta \) must be between 0 and \( \alpha \). The additive case that has been examined in previous papers is \( \delta = 0 \), which corresponds approximately to a rare dominant trait. The other extreme is the case \( \delta = \alpha \), which is approximately true for rare recessive traits. We include a more detailed discussion of the interpretation of \( \alpha \) and \( \delta \) for various modes of inheritance below.

We test for linkage by looking for extreme values of \( X_r \). One way to do this is to look for extreme values of a stochastic process that is an appropriate one-dimensional summary of \( X_r \). In our previous work, we examined the process \( X_{2,r} + X_{1,r} \), which is appropriate if one and two alleles IBD cannot be distinguished, and \( 2X_{2,r} + X_{1,r} \), which is the score statistic for an additive model \( (\delta = 0) \). Here we also consider \( X_{2,r} \), the score statistic when \( \delta = \alpha \). For each of these one-dimensional processes, our overall test statistic is the maximum value attained by the process over the length of the genome. The p-value associated with an observed peak of height \( b \) is the probability that the maximum exceeds \( b \) under the null hypothesis that there is no linked gene on the genome.

Feingold (1993) discussed appropriate theory for finding such p-values. For a large class of affected relative pair statistics, an approximation to the p-value is

\[
1 - \exp \left[ -\beta \binom{N}{b} p_0 (1 - p_0)^{N-b} (b - Np_0) \right],
\]

where the binomial probability \( \binom{N}{b} p_0 (1 - p_0)^{N-b} \) is the stationary probability that...
the process is in state $b$, $l$ is the total length of the portion of the genome being examined, and $\beta$ is a constant associated with the process. The case of $X_{2,t} + X_{1,n}$ for which (1) applies with the values $p_0 = 3/4$ and $\beta = 16\lambda/3$ was fully described in Feingold (1993). The case of $X_{2,t}$ is very similar: the appropriate parameters for equation (1) are $p_0 = 1/4$ and $\beta = 16\lambda/3$. The case of $2X_{2,t} + X_{1,t}$ can also be handled similarly by noting that a single sibling pair is equivalent to the sum of two independent half-sibling processes (under the null hypothesis of no linked gene on the chromosome), since the maternal and paternal meioses are independent. Thus $N$ sibling pairs can be treated exactly as $2N$ half-sibling pairs, a case that was covered in Feingold (1993). The approximate p-value is a special case of (1) with $p_0 = 1/2$, $\beta = 4\lambda$, and with $N$ replaced by $2N$.

The power of any of these tests can be written as

$$P(Y_r \geq b) + P(Y_r < b, Y_t \geq b \text{ for some } t \neq r),$$

where $r$ is the trait locus and $Y$ is the one-dimensional process of interest. This expression omits the probability of finding a significant peak on chromosomes other than the one containing the trait locus $r$, and should thus be interpreted as the power to find a peak on the true chromosome. The first term of the power expression is usually relatively easy to evaluate under the alternative hypothesis as a binomial probability or by a normal approximation with continuity correction. By itself it often provides a reasonable approximation to the power of the test; the more complicated second term usually adds about 5 - 10% to the total. Approximations for the second term are discussed in detail in the appendix.
**Gaussian Models**

In the case where \( N \) is large, following Feingold et al. (1993), we can analyze the approximately Gaussian process

\[
Z_t = (Z_0, t, Z_1, t, Z_2, t) = N^{-1/2} \left( X_{0, t} - N/4, X_{1, t} - N/2, X_{2, t} - N/4 \right). 
\]

The calculations for this approximating model are always simpler than for the Markov chain models, but may not be as accurate for small sample sizes and hence must be used with some care. (Sometimes it is useful to transform the process before applying Gaussian approximations.) The Gaussian versions of the one-dimensional processes we looked at for Markov chains are \( Z_{2, t} \) and \( Z_{2, t} + \frac{1}{2} Z_{1, t} \). It will be helpful to introduce the more general process \( T_t(c) = Z_{2, t} + cZ_{1, t} \). In addition to the values \( c = 0 \) and \( c = 1/2 \), which are useful for recessive and dominant modes of inheritance, respectively, we shall see below that the value \( c = 1/4 \) provides a useful compromise when the degree of dominance is unknown. Another interesting value is \( c = (1 - \kappa)/2(1 + \kappa) \), which gives the score statistic when \( \delta/\alpha = \kappa \). Usually this ratio will be unknown, so this statistic cannot be used in practice, but it provides an ideal against which to evaluate the performance of the other statistics.

If there is only one gene on the chromosome associated with the trait, the process \( Z_t \) has expected value at the trait locus of \( N^{1/2}[-\alpha/4, -\delta/2, (\alpha + 2\delta)/4] \). For a trait at locus \( r \), the expected value of \( T_t(c) \) can be written as \( \xi R_t(t - r) \), where

\[
\xi = N^{1/2}[(\alpha + 2\delta(1 - c))/4] 
\]

and

\[
R_t(x) = \frac{\alpha + \delta}{\alpha + 2\delta(1 - c)} e^{-4\delta|x|} - \frac{\delta(2c - 1)}{\alpha + 2\delta(1 - c)} e^{-\delta(2c - 1)\delta|x|}. 
\]
The function $R_1(x)$ can be approximated as $1 - \beta|x|$, when $|x|$ is close to zero, where

$\beta_1 = 4\lambda[\alpha + \delta(3 - 4c)][\alpha + 2\delta(1 - c)]$. The covariance of $T_i(c)$ and $T_j(c)$, where $s$ and $t$ are any two points on the same side of $r$, can be written as

$\sigma^2 R(t - s)$, where $\sigma^2 = [2 + (2c - 1)^2]/16$ and

$$R(x) = \frac{2}{2 + (2c - 1)^2}e^{-4\lambda|x|} + \frac{(2c - 1)^2}{2 + (2c - 1)^2}e^{-8\lambda|x|}.$$ 

The function $R(x)$ can be approximated as $1 - \beta|x|$, where

$\beta = 8\lambda[2(2c - 1)^2]/[2 + (2c - 1)^2]$. (This constant $\beta$ is the same as that used for the corresponding Markov chains.)

An approximation to the p-value (Feingold et al. 1993, Lander and Schork 1994) of the observed maximum value $a = \max \frac{T_i(c)}{\sigma}$ over $n$ chromosomes is

$$1 - \exp[-n(1 - \Phi(a)) - \beta l a \phi(a)],$$

(2)

where $\phi$ and $\Phi$ are the standard normal density and distribution, respectively, and $l$ is the total genetic length of the region of the genome being searched. The correspondence between $a$ in (2) and $b$ in (1) is $a = (b - NP_0)[NP_0(1 - NP_0)]^{1/2}$.

For a genome of 23 chromosomes and a total genetic length of 3000cM, the value $a = 4.1$ gives a p-value of about 0.05.

An approximation to the power (Feingold et al. 1993) is

$$1 - \Phi(a - \frac{\xi}{\sigma}) + \phi(a - \frac{\xi}{\sigma}) \left[2\frac{\beta}{\beta_1}\frac{\sigma}{\xi} - \left[\frac{\xi}{\sigma}(2\frac{\beta_1}{\beta} - 1) + a\right]^{-1}\right].$$
For sample size calculations we find numerically the value of $\xi/\sigma$ that gives the desired level of power, set that equal to the algebraic expression for $\xi/\sigma$ as a function of $N$, $\alpha$, and $\delta$, and solve for $N$.

**Interpretation of $\alpha$ and $\delta$**

For any trait, the parameter $\alpha$ must lie between zero and one, and the parameter $\delta$ must lie between 0 and $\alpha$. Suarez et al. (1978) calculated the values of $\alpha$ and $\delta$ for a single-gene, two allele trait as

$$\alpha = \frac{(V_A/2 + V_D/4)}{(K^2 + V_A/2 + V_D/4)} \quad \text{and} \quad \delta = \frac{(V_D/4)}{(K^2 + V_A/2 + V_D/4)},$$

where $K$ is the population prevalence of the trait and $V_A$ and $V_D$ are the additive and dominance variances of the penetrances. It can easily be shown that the two-allele assumption is not necessary for these formulas to hold. Risch (1990b) did similar calculations in terms of relative risks to different kinds of relatives, and got

$$\alpha = \frac{(\lambda_S - 1)/\lambda_S}{\lambda_S}, \quad \delta = \frac{(\lambda_S - \lambda_O)/\lambda_S}{\lambda_S}, \quad (3)$$

where $\lambda_S$ is the relative risk to siblings of affecteds and $\lambda_O$ is the relative risk to offspring of affecteds. These latter representations have the advantage that in principle they can be estimated directly from pedigree data; their disadvantage is that they do not generalize simply and naturally to pedigrees containing more than two affecteds.

The above representations are appropriate in the presence of phenocopies, but they are not correct when there are other major genetic loci involved with the
trait. Risch (1990a) (cf. also Dupuis et al. 1995) introduces models for complex
disease involving more than one susceptibility locus. A particularly interesting
special case that serves as an approximate model for locus heterogeneity of a rare
trait is the additive model, where the total penetrance is assumed to be the sum of
penetrances of susceptibility alleles at unlinked loci in linkage equilibrium. In this
case there are parameters \( \alpha_i, \delta_i \) associated with the \( i \)th trait susceptibility locus,
which can be expressed in terms of locus specific additive and dominance variances
of the penetrances:

\[
\alpha_i = \frac{V_{Ai}/2 + V_{Di}/4}{K^2 + \sum_j [V_{Aj}/2 + V_{Dj}/4]} \quad \delta_i = \frac{V_{Di}/4}{K^2 + \sum_j [V_{Aj}/2 + V_{Dj}/4]}
\]

Equation (3) continues to hold with \( \alpha \) replaced by \( \Sigma \alpha_i \) and \( \delta \) replaced by \( \Sigma \delta_i \).
The degree of dominance can vary from one locus to the next.

The parameter \( \delta \) contains, roughly speaking, the information about how
recessive or dominant the locus is. For a relatively rare single-locus trait,
dominance corresponds approximately to the additive case, where \( \delta = 0, V_D = 0, \)
and \( \lambda_S = \lambda_O \). The other extreme is \( \delta = \alpha \), which occurs when \( V_A << V_D \) and
\( \lambda_0 << \lambda_S \). This is a good approximation to many recessive traits. We show in the
next section that choosing the best statistic for finding a disease locus depends on
“how recessive” the locus is, as expressed in terms of the ratio \( \delta/\alpha \). Using the
representations above, we see that for locus \( i \), the ratio \( \delta_i/\alpha_i \) equals
\( V_{Di}/(2V_{Ai} + V_{Di}) \), even in the presence of phenocopies or locus heterogeneity.
Examination of a two-allele model for locus $i$ gives further insight into exactly what characteristics give the large values of $\delta / \alpha_i$ that make a locus "recessive" for the purposes of choosing the best statistic. In a two-allele model, we can assign the three possible genotypes to have probabilities $q^2$, $2pq$, and $p^2$, where $p$ is the frequency of the susceptibility allele, and the penetrances can be written as $f_0$, $f_0 + f + d$, and $f_0 + 2f$. The parameters $d$ and $f$ describe the contribution from the locus being investigated. Note that $d = f$ for a dominant trait, $d = 0$ for an additive trait, and $d = -f$ for a recessive trait, so that we are primarily interested in negative values of $d$. The prevalence, $K$, of the trait is $f_0 + 2pqd + 2pf$, the dominance variance is $V_{DI} = 4p^2q^2d^2$ (Kempthorne 1957), and the additive variance is $V_{AI} = 2pq[f + d(q - p)]^2$. Then

$$\frac{\delta / \alpha_i}{\alpha_i} = \frac{pq(d/f)^2}{pq(d/f)^2 + [1 + (d/f)(q - p)]^2}.$$ 

Note that $f_0$ does not play a role in determining the size of this ratio, while $d$ and $f$ affect the ratio only through the value $d/f$. If the locus is completely recessive in the usual sense ($-d/f = 1$), we get $\delta / \alpha_i = (1 - p)/(1 + 3p)$, which is greater than $1/2$ as long as $p < 0.2$. Thus a fully recessive locus is best found by a "recessive statistic" designed for large $\delta / \alpha$ as long as the susceptibility allele is moderately rare. A common allele that slightly increases the likelihood of disease, even if that increase acts recessively, will not give a large value of $\delta / \alpha$. If the locus is not fully recessive, i.e., if $-d/f < 1$, the picture changes markedly, and values of $\delta / \alpha$ can be much lower. The ratio $\delta / \alpha$ takes a maximum value of
\((d/l)^2/[4 - 3(d/l)^2]\), achieved at allele frequency \(p = (1 + d/l)/2\). Thus for \(d/l = -0.9\), the maximum value of \(\delta/\alpha\) is 0.52, and will be lower if the allele frequency is far from the maximizing value of 0.05. For \(d/l = -0.8\), the maximum value of \(\delta/\alpha\) is only 0.31. Two examples that are illustrative here are chronic obstructive pulmonary disease (COPD) and late-onset Alzheimer’s disease (AD).

In the case of COPD, the PiZ allele is only one of many possible causes (Khoury et al. 1993), but since it is relatively rare and acts, apparently, completely recessively, it has a high value of \(\delta/\alpha_i\). Khoury et al. give values of 0.05 for the trait frequency, 0.02 for the allele frequency, and 20 for the risk ratio to homozygotes, yielding \(\delta/\alpha_i = 0.92\). For AD and the APOE-ε4 allele, Tsai et al. (1994) give odds ratios of 3.6 for heterozygotes and 18.3 for homozygotes, and an allele frequency of 0.13. Depending on one’s assumption about the overall trait frequency, this gives \(-d/l\) of at most 0.7, and thus \(\delta/\alpha_i\) of at most 0.19.

The parameter \(\alpha\) essentially contains the information about how important the locus is. Sample sizes are heavily dependent on \(\alpha\) whatever statistic is used. If there is only one locus, \(\alpha = 1\) when the trait-causing allele is rare and the level of phenocopies is low. In the presence of phenocopies, \(\alpha\) is smaller, approaching zero as the proportion of cases attributable to phenocopies approaches one. Under heterogeneity, the preceding remark applies to \(\Sigma \alpha_i\); each individual \(\alpha_i\) will be smaller. For the COPD example above, \(\alpha_i = 0.035\), because, though the allele is relatively rare, the level of phenocopies/heterogeneity is very high. This value of \(\alpha_i\) would require a sample size on the order of 7000 sibling pairs to detect with 90% probability. Of course, if a trait is carefully defined to eliminate heterogeneity and phenocopies as much as possible, \(\alpha_i\) increases and the sample size is smaller.
Given the population prevalence of Alzheimer's under varying definitions of the disease, and given the data of Tsai et al., it seems likely that a value of $\alpha_i = 0.5$ could be achieved, requiring a sample size on the order of 150 sibling pairs for a hypothetical attempt to map the APOE locus with 90% power.

**Comparison of Sibling Pair Statistics**

For any given value of $\kappa = \delta/\alpha$, sample sizes for the statistic $T_i(c) = Z_{2,1} + cZ_{1,1}$ are directly proportional to $1/\alpha^2$, so the relative sample sizes for different values of $c$ can be compared as a function of $\kappa$ without taking $\alpha$ into account. For a false positive error rate of 0.05, power of 0.90, and a genome length of 3000 cM, sample sizes for the ideal value $c = (1 - \kappa)/2(1 + \kappa)$ range from 8.5/\alpha^2 for $\delta = \alpha$ to 50.1/\alpha^2 for $\delta = 0$. Figure 1 shows the performance of the statistic $T_i(c)$ for $c = 1/2, 1/4$, and 0, expressed in terms of the percentage increase needed in the sample size compared to the ideal value of $c$. The horizontal axis of figure 1 is $\kappa = \delta/\alpha$, which varies from 0 to 1. As expected, the "recessive" statistic $T_i(0)$ performs very well when $\kappa$ is close to 1, and the "dominant" statistic $T_i(1/2)$ performs very well when $\kappa$ is close to 0. The compromise statistic $T_i(1/4)$, which to achieve 90% power requires a sample size no more than about 12% larger than the optimum over the entire range of $\kappa$, would seem to be a reasonable choice when the mode of inheritance is unknown. The simpler $T_i(0)$ performs very well when the trait is at least moderately recessive ($\kappa$ is at least 1/2).

The simple comparisons of figure 1 are based on the Gaussian approximations, which to some extent make $T_i(0)$ look better than it actually is
when the sample size is small to moderate. The reasons are that on unlinked chromosomes the distribution of $T_i(0)$ is skewed, so the correct threshold is slightly larger than the Gaussian approximation suggests, and at a trait locus the true variance of $T_i(0)$ is larger than the variance used in the Gaussian approximation, so the power is actually less than the approximation suggests. This approximation is, nevertheless, useful, because it gives us a simple qualitative picture based on the single parameter $\kappa$ instead of a more accurate but more difficult to interpret description involving $N$, $\alpha$, and $\delta$. For the numerical calculations below we have used the more accurate Markov chain approximations.

These comparisons may at first glance appear inconsistent with the results of Knapp et al. (1994a,b), who show that the statistic $T_i(1/2)$ is in a certain sense optimal for detecting linkage of a recessive trait. However, they assume a monogenic recessive trait without phenocopies. Under these assumptions, when $p$ is small only small sample sizes are required, and both $T_i(0)$ and $T_i(1/2)$ are about equally efficient; when $p$ is not so small, $\kappa$ is not close to one, so $T_i(1/2)$ can be more efficient than $T_i(0)$. For example, when $p = 0.1$, already rather large for this mode of inheritance, fifteen pairs of sibs are required to achieve 90% power if we use $T_i(1/2)$ and sixteen pairs if we use $T_i(0)$. For $p = 0.25$ the required sample sizes become 32 and 37, respectively. A case not covered by these assumptions is a rare heterogeneous trait. In this case it is quite possible to have $\kappa$ about equal to one yet have $\alpha$ relatively small, so a large sample size is required. For example, if four loci make equal contributions to a rare trait, so $\alpha = \delta = 0.25$, then to obtain 90% power one would need about 204 observations when using
$T_A(1/2)$ and about 148 when using $T_A(0)$. These considerations would also apply to detection of linkage at a locus where the trait allele is rare, even though overall the trait might be relatively common.

Two other statistics that have previously been suggested to deal with the case that the mode of inheritance is unknown are the likelihood ratio, or "possible triangle," test (Holmans 1993) and $Z_{max} = \max(T_A(0), T_A(1/2))$ (Schaid and Nick 1990). The performance of these statistics is similar to that of $T_A(1/4)$. A discussion is contained in the Appendix.

An Example: Type 1 Diabetes

An illustrative example from recent literature is the use of sibling pairs to detect linkage for type 1 diabetes (Davies et al. 1994; Hashimoto et al. 1994). Type 1 diabetes appears to be a heterogeneous disease with a mode of inheritance intermediate between additive and recessive, and a previously known finding of linkage to HLA on chromosome 6. Thus it is a very appropriate application of the theory developed in this paper, as well as of the extension of that theory to conditional search for additional loci beyond the HLA locus. Both Davies et al. and Hashimoto et al. employ a genome wide search, and both confirm earlier findings of linkage to HLA. According to Kruglyak and Lander (1995) the data of Davies et al. yield estimated values at HLA of $\alpha_1 = .61$ and $\delta_1 = .36$. For these parameter values a sample size of about 53 is required for 90% power for any of the statistics $T_A(c)$.

Remark. Allele sharing data at a specific locus, which may be one of several (unlinked) trait susceptibility loci, allow one to estimate directly the probability of
identity by descent and hence the parameters $\alpha_1$ and $\delta_1$ associated with that locus without assumptions about the mode of inheritance. Kruglyak and Lander give estimates of $\lambda_2$ and $\lambda_0$, which can be obtained from estimates of $\alpha_1$ and $\delta_1$ via equation (3), provided we make assumptions about the mode of inheritance.

Equation (3) presupposes a monogenic trait, or at least a trait where gene interaction is multiplicative, in which case $\lambda_2$ and $\lambda_0$ are locus specific recurrence risks (cf. Risch 1990a).

To search for additional trait susceptibility loci Davies et al. use a form of what Dupuis et al. (1995) call conditional search, where one conditions on the observed data at the locus already found to be linked in order to magnify the signal at other putative trait susceptibility loci. Davies et al. stratify their data into pairs that are identical by descent on both their paternal and maternal chromosomes at HLA and those identical by descent on 0 or 1 chromosome; and they search in each stratum separately the rest of the genome for additional trait susceptibility loci.

Dupuis et al. stratify the data according to the number of alleles shared identical by descent at a linked locus, HLA in this case, and use an optimally weighted combination of statistics from the three strata (pairs identical by descent on 0, 1, or both chromosomes at HLA). It is interesting to apply their methods in the context of this paper to see how much efficiency might be gained by using an optimal statistic in place of the ad hoc methods of Davies et al.

Let $X_{i,j,t}$ denote the number of pairs sharing $i = 0, 1, \text{or } 2$ alleles identical by descent at the HLA locus and $j = 0, 1, \text{or } 2$ alleles identical by descent at a putative trait susceptibility locus $t$ on a different chromosome. Also let
\[ X_{i,t} = \sum_j X_{i,j,t} \] be the number of pairs sharing \( i \) alleles identical by descent at HLA, and \( X_{i,j,t} \) be the number of pairs sharing \( j \) alleles identical by descent at locus \( t \). Let \( Z_{i,j,t} = (X_{i,j,t} - 2^{(2-\alpha)}X_{i,t})/N^{1/2} \) where \( 1/j \) equals 1 if \( j = 1 \) and equals 0 if \( j \) equals 0 or 2. Let \( T_{i,t}(c) = Z_{i,2,t} + cZ_{i,1,t} \). The theory developed by Dupuis et al. applied to our problem indicates that if the values \( \alpha_1 \) and \( \delta_1 \) associated with HLA were known, the optimal combination of the \( T_{i,t}(c) \) would be

\[
(1 + \alpha_1 + 2\delta_1)^{-1}T_{2,t}(c) + (1 - \delta_1)^{-1}T_{1,t}(c) + (1 - \alpha_1)^{-1}T_{0,t}(c)
\]  \hspace{1cm} (4)

which can be normalized to have unit variance at unlinked loci by dividing by

\[
\left\{ \frac{4(1 + \alpha_1 + 2\delta_1)}{[2(1 - \delta_1)]^{-1} + [4(1 - \alpha_1)]^{-1}} \right\}^{1/2}
\]  \hspace{1cm} (5)

Since the parameters \( \alpha_1 \) and \( \delta_1 \) are in fact unknown, the conditional search statistic involves substituting the maximum likelihood estimates \( \hat{\alpha}_1 \) and \( \hat{\delta}_1 \) into (4) and (5) and using the ratio thus obtained of (4) to (5) to test for additional linkages. The efficiency of this statistic compared to unconditional search is the square of (5). If we use the estimates provided by Kruglyak and Lander, we see that conditional search is about 50% more efficient than unconditional search.

It is clear intuitively that for a disease with a recessive like mode of inheritance at the already detected locus (HLA in this case) the greatest gain from conditioning will come from those pairs that are identical by descent on 0 or 1 chromosome. This is also apparent from (4), where the greatest weights are
associated with those pairs. However, many pairs at the conditioned locus, which
is presumed to be linked, will be identical by descent on both chromosomes. This
means that those pairs that are individually most informative are also the fewest in
number, while those that are least informative are the most plentiful. As a
consequence one gets more information from the weighted sum in (4) than from the
individual terms. If $\delta_1 = \alpha_1$, and especially if these parameters are large, the
procedure actually used by Davies et al. of lumping together pairs identical by
descent on 0 or 1 chromosome would be approximately optimal. Its relative
efficiency compared to unconditional search is $(9/16)[(1 - \delta_1)/2 + (1 - \alpha_1)/4]$,
which for the values $\alpha_1 = 0.61$, $\delta_1 = 0.36$ equals 1.35. Hence this conditional
search is 35% more efficient than unconditional search, but is about 11% less
efficient than the conditional search of Dupuis et al.

**Efficiency of Sibling Pairs vs. Cousin Pairs**

Several authors (e.g. Risch 1990b, Elston 1992, Feingold et al. 1993) have
noted that for additive traits, defined by the condition $\lambda_S = \lambda_O$, pairs of more distant
relatives are more efficient than sibling pairs when $\lambda_S = \lambda_O$ is large. For
completely recessive traits, where $\lambda_S >> \lambda_O$, clearly unilineal relative pairs are not
useful. But it is interesting to see how they compare to sibling pairs when a trait is
partially recessive. Figure 2 shows the relative efficiency of sibling pairs to first
cousin pairs. For additive traits as a function of $\lambda_S$, sibling pairs are more powerful
than cousin pairs only when $\lambda_S = \lambda_O$ is small (approximately $\lambda_S < 2$). The other
case we consider is $\lambda_O = (1 + \lambda_S)/2$, which by (3) would be equivalent to $\delta = \alpha/2$
for a monogenic trait. The relative efficiencies given in the figure also apply for a
heterogeneous trait provided $\delta_i = \alpha_i/2$ at the locus under consideration. (This means the degree of recessiveness observed at the pedigree level for the trait as a whole is the same as at the particular locus. The situation would be quite different if the overall intermediate mode of inheritance resulted from some loci behaving dominantly and others behaving recessively.) For this partially recessive case, sibling pairs are more efficient than cousin pairs up to about $\lambda_S = 13$, and are much more efficient for small $\lambda_S$.

**Inbred Individuals**

Inbred individuals have only recently been used for mapping studies (e.g. Thomas et al. 1995, Gschwend et al. 1995), though the potential for doing so has long been recognized (Smith 1953, Lander and Botstein 1987). The statistical analysis of HBD data for inbred individuals is essentially the same as the analysis of IBD data for affected pairs. A stochastic process can be constructed to describe each type of consanguineous mating, and p-values can be calculated by the same methods described above. A practical computational method for carrying on such analyses is given by Kruglyak, Daly and Lander (1995).

**Markov Chain Models**

For the $i$th individual, let $Y_{t(i)}$ equal one when there is HBD and zero when there is not. Unfortunately, $Y_{t(i)}$ is not a Markov chain, but it can be
described as a function of an underlying, unseen Markov chain. The underlying Markov chains are different for each type of consanguineous mating (e.g. parent/child, first cousin, second cousin), and are described below. The \( Y_t^{(i)} \)s for a given mating type can be added together, as with sibling pairs, to give a process \( Y_t \) that records the number of individuals who are HBD at each marker locus. As with sibling pairs, if there is no linked gene on the chromosome, \( Y_t \) starts in its stationary distribution and remains there for the length of the chromosome. The stationary distribution is binomial \( (N, F) \), where \( F \) is the coefficient of inbreeding. If there is a linked gene at the locus \( r \), the distribution at \( r \) is binomial \( [N, F + \gamma(1 - F)] \). The parameter \( \gamma \) takes values between 0 and 1, indicating the amount of excess HBD. It actually depends on \( F \), and is discussed further below.

In order to approximate p-values and power, we need to look at the Markov chains underlying \( Y_t \) for each type of consanguineous mating. Figure 3 shows a sample pattern of HBD on one pair of chromosomes in the offspring of a parent/child mating. It can be seen from the diagram that the offspring is HBD at a given locus only if (1) the offspring and its maternal grandmother do not share IBD at that locus, and (2) the offspring and its mother do share IBD on their respective paternally derived chromosomes. This information allows us to construct an underlying Markov chain of which \( Y_t^{(i)} \) is a function. Let \( X_t^{(i)} = (X_t^{(ia)}, X_t^{(ib)}) \), where \( X_t^{(ia)} \) is the IBD process between the paternally derived chromosomes of the mother and offspring that equals one if those two chromosomes are IBD and zero if they are not, and \( X_t^{(ib)} \) is the IBD process between the inbred offspring and its maternal grandmother. The processes \( X_t^{(ia)} \) and \( X_t^{(ib)} \) are independent Markov chains, and are
described in more detail in Feingold (1993) as half-sibling and grandparent/grandchild processes, respectively. The process we actually observe, $Y_t^{(i)}$, equals one only when $X_t^{(i)} = (1, 0)$. If we are looking at $N$ offspring of parent/child matings, $Y_t = \sum Y_t^{(i)}$ is the total number of offspring HBD at the locus $t$. The approximate p-value, by the methods of Feingold (1993), is given by equation (1) with $p_0 = 1/4$ and $\beta = 4\lambda$. For a half-sibling mating, a similar analysis shows that (1) applies with $p_0 = 1/8$ and $\beta = 32\lambda/7$. The other cases we have examined do not require the description of new underlying chains, because they are identical to affected-pair processes described in Feingold (1993). For example, the HBD process for a child of siblings is exactly the same as the IBD process for first cousins, because the child of siblings can be regarded as its own first cousin. The p-values can be approximated by (1) with parameters as given in table 1.

Power and sample size calculations can be done in the same manner as for sibling pairs, using the fact that the distribution of $Y_t$ is binomial $[N, F + \gamma(1 - F)]$.

**Gaussian Models**

The Gaussian models for inbred individuals can be defined just as they were for sibling pairs: $Z_t = N^{-1/2}(Y_t - NF)$, which has expected value at the trait locus of $\xi = \sqrt{N} \gamma(1 - F)$. Under the hypothesis of no linkage, the covariance of $Z_t$ and $Z_j$ is $F(1 - F)R(t - s)$. Formulas for $R(x)$ are given in table 2, and the corresponding values of $\beta$ are given in table 1. However, since inbred individuals tend to be most useful for rare traits, which often require only small sample sizes to detect linkage, and because $F$ is often small, $F = 1/16$ for a child of first cousins or $F = 1/64$ for
a child of second cousins, one should not expect a Gaussian process approximation to the distribution of \( \bar{Z}_r \) under the hypothesis of no linkage to be accurate. However, this approximation can be vastly improved by transforming the data first. A particularly useful transformation is to use the likelihood ratio statistic instead of the score statistic. To that end we let

\[
Z_t = \pm 2^{1/2}(Y_t \log(Y_t/NF) + (N - Y_t)\log[(N - Y_t)/N(1 - F)])^{1/2}
\]

The plus or minus sign is to be interpreted as plus when \( Y_t \geq NF \) and minus otherwise. This statistic is the signed square root of twice the log likelihood ratio statistic and for each \( t \) is approximately normally distributed with mean 0 and variance 1 under the null hypothesis of no linkage. It is a monotonic function of \( \bar{Z}_r \), so in principle both statistics yield the same test. But one can expect the Gaussian approximation \((2)\) to be more accurate when applied to \( Z_t \) than to \( \bar{Z}_r \).

The covariance function for \( Z_t \) is the same function \( R(t) \) as for \( \bar{Z}_r[F(1 - F)]^{1/2} \). The expected value of \( Z_t \) on a linked chromosome having a single trait locus at \( r \) is approximately \( \xi R(t - r) \), where

\[
\xi = N^{1/2}2^{1/2}[(F + \gamma(1 - F)]\log[1 + \gamma(1 - F)F^{-1}] + (1 - F)(1 - \gamma)\log(1 - \gamma)]^{1/2}.
\]

\( P \)-values, power, and sample sizes can be approximated by the methods used for sibling pairs. Note that there is not a distinct function \( R_1(x) \) in the case of inbred individuals, and thus \( \beta_1 = \beta \).
Remark. It is not generally true that the likelihood ratio statistic is a function of the score statistic. In particular for sibling pairs the two statistics are genuinely different. According to general statistical theory (Cox and Hinkley, 1974, Chapter 9), in large samples the two statistics behave about the same under the null hypothesis and for small nonnull values of the parameter (γ in the case of inbred individuals, α, δ in the case of siblings). Thus in cases that require large sample sizes there is no strong reason to prefer one statistic to the other. However, for large departures from the null hypothesis, i.e., when detection is comparatively easy and relatively small sample sizes are required, the two can perform differently, and there does not seem to be any simple, general way to know which is more powerful. In view of its relative simplicity, we have focused on the score statistic in our previous work.

Interpretation of γ

In order to compare the utility of different types of inbred individuals, we need to know how the parameter γ varies with F. A general one-locus n-allele model sets the penetrance of an individual with alleles i and j equal to $K + f_i + f_j + d_{ij}$ (Kempthorne 1957), where K is the prevalence of the trait in the outbred population, $f_i$ is the net additive effect of allele i, and $d_{ij}$ is the net effect of the interaction between alleles i and j. Writing the frequency of allele i as $p_i$, we have $\sum_i p_i f_i = 0$ and $\sum_i p_i d_{ij} = 0$. Then the probability that an inbred individual is HBD given that he or she is affected is
\[ F + \gamma(1 - F) = \frac{P(HBD)P(\text{affected} \mid HBD)}{P(\text{affected})} = \frac{F(K + \sum p_d_{ii})}{K + F\sum p_d_{ii}}. \]

so

\[ \gamma = \frac{F\sum p_d_{ii}}{K + F\sum p_d_{ii}}. \]

Thus the relationship between values of \( \gamma \) for different types of inbreeding depends on \( F \), on \( K \), and on \( \sum p_d_{ii} \), making it fairly complicated. We gain some simplification under the two-allele model introduced in the discussion of \( \alpha \) and \( \delta \), where \( \sum p_d_{ii} = -2pqd \). The above formula for \( P(HBD|\text{affected}) \) is correct in the presence of both phenocopies and of heterogeneity with other loci that behave additively. If there is heterogeneity with another recessive locus, we have

\[ F + \gamma_1(1 - F) = P(\text{HBD at locus 1|affected}) = \frac{F(K + \sum p_d_{ii} + \sum p_d'_{ii})}{K + F\sum p_d_{ii} + F\sum p_d'_{ii}}, \]

and thus

\[ \gamma_1 = \frac{F\sum p_d_{ii}}{K + F\sum p_d_{ii} + F\sum p_d'_{ii}}, \]

where the primes indicate the values for locus 2.
Values of $\gamma$ must be between 0 and 1. For a given trait, $\gamma$ is highest for offspring of close relatives, but that does not necessarily imply greater mapping power for closer relatives. However, for a given inbreeding type, traits with higher values of $\gamma$ will be more easily mapped than those with lower values of $\gamma$. Under the one-locus, two-allele model, $\gamma$ is high when the trait allele is rare and the level of phenocopies is low. For example, for COPD, $\gamma$ for first cousin offspring is only 0.02, because the level of phenocopies is so high. A trait with the same allele frequency in which only half the cases are phenocopies would give $\gamma = 0.61$, and with no phenocopies would give $\gamma = 0.75$. Under heterogeneity, values of $\gamma_i$ decrease as the number of loci increases, with $\Sigma \gamma_i$ never exceeding one. With very few phenocopies and $n$ equally contributing recessive loci each with a very rare trait allele, $\gamma_i = 1/n$, and the probability of HBD at each trait locus is approximately $[1 + F(n - 1)]/n$.

In order to compare the mapping power of inbred individuals to that of sibling pairs, it is necessary to relate the parameters $\gamma$, $\alpha$, and $\delta$. For the one-locus, two-allele model, the probability that an inbred individual is affected given that he or she is HBD at the trait locus is $qf_0 + p(f_0 + 2f) = f_0 + 2pf = K + \sqrt{V_D}$ for $d \leq 0$. Since $V_D = 4p^2q^2d^2 = 4K^2\delta(1 - \alpha)$, we have

$$P(\text{HBD} \mid \text{affected}) = \frac{F(K + \sqrt{V_D})}{F(K + \sqrt{V_D}) + (1 - F)K}$$
\[ F + (1 - F) \left( \frac{2F\sqrt{\frac{\delta}{1 - \alpha}}}{1 + 2F\sqrt{\frac{\delta}{1 - \alpha}}} \right). \]

For a heterogeneous trait, modeled by assuming additivity of penetrances between, for example, two loci in linkage equilibrium, the probability of HBD at both trait loci is \( F^2 + F(1 - F)(\gamma_1 + \gamma_2) \); the probability of HBD at the first but not the second trait locus is \( F(1 - F)[1 + (1 - F)\gamma_1/F - \gamma_2] \), etc. When both trait loci are diallelic,

\[ \gamma_i = \frac{2F[\delta_i(1 - \alpha_1 - \alpha_2)]^{1/2}}{1 + 2F[\delta_1^{1/2} + \delta_2^{1/2}]/(1 - \alpha_1 - \alpha_2)^{1/2}}. \]

Table 3 shows values of \( \alpha, \delta, \gamma_1 \) (first cousins), and \( \gamma_2 \) (second cousins) for a number of different variations on the two-allele model. For a single locus with homozygote penetrance \( (\delta_0 + 2f) \) equal to one and \(-dlf = 1\), parameter values vary with the phenocopy level and the trait frequency. For a given trait frequency, the presence of phenocopies decreases \( \alpha \) slightly, increases \( \delta \) slightly, and decreases \( \gamma_1 \) and \( \gamma_2 \) moderately, indicating that phenocopies damage the usefulness of inbred individuals more than that of sibling pairs. Rarer traits yield higher values of all parameters. If the homozygote penetrance is only 0.5, parameter values are somewhat lowered, but not dramatically so. However, if \(-dlf = 0.9\), values of \( \delta, \gamma_1, \) and \( \gamma_2 \) are quite low, indicating that “recessive trait” mapping strategies will be less useful. As noted earlier, the value of \( \delta \) in this case varies non-monotonically with the allele frequency, and thus can be somewhat unpredictable. With multiple loci, the picture is similar. (For the multiple-locus
calculations, the loci are assumed to be equally contributing, unlinked, and in linkage equilibrium.)

Comparison of Different Types of Inbred Individuals and Sibling Pairs

Table 4 gives some sample sizes necessary for 90% power to detect linkage using sibling pairs and offspring of first and second cousins for selected combinations of $\alpha$ and $\delta$. Each entry corresponds to an entry in table 3. For sibling pairs the sample size depends only on the given values of $\alpha$ and $\delta$, not on the underlying genetic parameters shown in table 3 or on the two-allele model. The related values of $\gamma_1$ and $\gamma_2$ are specific to the two-allele model and the number of loci (assumed to be equally-contributing). Since table 4 involves only large values of $\gamma$, the sibling sample sizes are computed using the statistic $T_1(0)$. The relation between $(\alpha, \delta)$ and $(\gamma_1, \gamma_2)$, hence the relative value of sibling pairs and inbred individuals, depends on the number of loci. For a single locus trait offspring of closer relatives are more powerful unless $\alpha$ is very close to one (few phenocopies and a very rare allele). Lander and Botstein (1987) obtained similar results, although they focused on marker heterozygosity and distance between markers rather than on phenocopies and heterogeneity as factors making the detection of linkage difficult. Except again for $\alpha$ near one, sibling pairs are substantially more powerful than inbred individuals. However, when $\alpha$ is close to one both siblings and inbred individuals require only small sample sizes, so the question of efficiency seems less important.
The situation is quite different for a rare heterogeneous trait with few phenocopies, e.g., xeroderma pigmentosum of Fanconi anaemia, because inbred individuals become more efficient relative to sibling pairs as the number of loci increases. In addition, if $\delta_i$ is about equal to $\alpha_i$, offspring of more distant relatives can be more powerful, with their advantage again increasing with the number of loci. In interpreting the sample sizes for heterogeneous traits in table 4, one must keep in mind that the sample size is that required to detect linkage to a given trait susceptibility locus and that all loci are assumed to contribute equally to the trait. The results would be qualitatively similar, although quantitatively different if a small number of loci account for a comparatively large part of the incidence of the trait, or if our requirement were different, e.g., to detect at least one of the trait loci or to detect all trait loci. The assumption that all trait loci contribute equally is a convenient conceptual device to simplify an intrinsically complex situation, but it is unlikely to be even approximately satisfied when several loci are involved. To explore the full range of possibilities there does not seem to be any substitute for performing a large number of computations under a variety of assumptions.

For example, Gschwend et al. (1995) scan the genome to detect linkage for Fanconi anaemia, which through complementation analysis is known to be heterogeneous and to involve at least five loci. Because of the rarity of the disease inbreds can be expected to be more powerful than sib pairs and offspring of second cousins more powerful than offspring of first cousins. Gschwend et al.'s sample of inbred affecteds, which was selected not to contain families segregating at a
previously linked trait locus on chromosome 9 (Strathdee et al. 1992), consisted of 23 pedigrees that involved primarily offspring of first cousin matings. Assuming for the sake of discussion that after excluding the chromosome 9 locus there are exactly four equally contributing loci and that the sample consists of exactly 23 offspring of first cousin matings, we find that the power to detect linkage to any given locus would be about 0.45, while the power to detect at least one of the four loci would be roughly \(1 - (1 - 0.45)^4\), or about 0.9. In fact, according to Gschwend et al. (1995), one locus appears responsible for a substantial majority of the total incidence, so the power to detect that locus is large, while the power to detect the other loci is very small.

The genome search whose properties are described in table 4 is predicated on the hypothesis of a single trait locus, although it will, of course, detect multiple loci if the observed statistic exceeds the threshold \(b\) at several places along the genome. It is also possible to design procedures specifically to search for multiple loci simultaneously (Lander and Botstein 1986, Dupuis et al. 1995) or sequentially (Dupuis et al. 1995). These require that one be prepared to make assumptions about the number of major loci and their mode of interaction. The resulting procedures can be considerably more powerful than single locus search when those assumptions are correct. See Gschwend et al. (1995) for an application of simultaneous search to Fanconi anaemia.

Population Considerations

The comparative usefulness of outbred siblings and inbred individuals for gene mapping also depends to some extent on their frequencies within the
population. For a monogenic recessive disease with a negligible level of phenocopies, caused by homozygosity of a rare allele of frequency $p$, the probability that both outbred parents of a pair of siblings are carriers is approximately $(2p)^2$. The conditional probability that both siblings are affected is $1/16$, so the probability that a randomly selected pair of outbred siblings have the disease is approximately $p^2/4$. Similarly, the probability that one or the other of the common grandparents of a pair of cousins is a carrier of the allele is approximately $4p$. The conditional probability that the cousins are both carriers is approximately $1/16$, and the conditional probability they would both pass this allele to their child is $1/4$. Hence if $\pi_C$ denotes the frequency of first cousin matings, the probability a child is the affected product of a first cousin mating is approximately equal to $p\pi_C/16$. Roughly speaking then, affected children of first cousin matings will be sufficiently common to be useful if the frequency of first cousin matings is large compared to the allele frequency $p$.

The task of finding affected inbred individuals for linkage analysis is presumably facilitated by concentrating on isolated or inbreeding populations. Thus in many cases in which both sibling pairs and inbred individuals are likely to be available, the inbred individuals will come from a distinct subpopulation that is not directly comparable to the outbred population. This does not mean that the two datasets cannot be used together to map the trait, but it does mean that it is difficult to quantify a priori the relative amounts of information in the two. Moreover, even if only inbred individuals are to be used, the theory developed here refers to an inbred individual in an otherwise outbred population. For an inbred individual in an
inbred population the probability of HBD at an arbitrary locus would be, in
Wright's notation (cf. Crow and Kimura 1970, p. 106), $F_{IT} = F_{IS} + (1 - F_{IS})F_{ST}$.
Here $F_{IS}$ is the coefficient of inbreeding due to the immediate family relation, which
we have been calling $F$ (1/16 for a child of first cousins), while $F_{ST}$ is the
coefficient of background inbreeding of the population. To consider an admittedly
extreme example, for the Hutterite population the value of $F_{ST}$ has been estimated to
be 0.04 (cf. Crow and Kimura 1970, p. 106). In such a population the probability
of HBD in a child of first cousins is 0.1, or about 50% larger than the 0.0625 of an
outbred population. To have a valid test of the null hypothesis of no linkage, one
should in principle use this corrected HBD probability.

Bodmer and Cavalli-Sforza (1976, p. 371) give estimated inbreeding
coefficients for a number of subgroups in different parts of the world. Their
estimates are often about 0.001, which are small enough to be ignored. In a few
cases of isolated populations they are about 0.01, which probably can be ignored
when dealing with offspring of first cousins, but might cause problems if one’s
data involved largely second cousins, where $F = 1/64 = 0.016$, so $F_{IT} = 0.026$.

**Discussion**

We have presented Markov chain and Gaussian models that allow us to
explore strategies for using sibling pairs and inbred individuals to map recessive or
partially recessive traits. Our measure of the degree of recessiveness is our
parameter $\kappa = \delta / \alpha$, which varies between 0 and 1, and is large when a double dose of a relatively rare allele causes a large increase in liability, especially if there is no heterozygote liability. For sibling pairs, we find the statistic $T/(1/4)$ to be quite efficient over a wide range of values of $\kappa$. This is also true of the other “compromise statistics” proposed by Holmans (1993) and by Schaid and Nick (1990). We also note that the previously known advantage of more distant relative pairs over sibling pairs for mapping additive traits does not hold for partially recessive traits in most cases. For recessive traits and inbred individuals, we find that offspring of closer relatives are more powerful in most cases, but that if the level of phenocopies is low, the allele is rare, and the trait is completely recessive, offspring of more distant relatives are better, with the advantage increasing as the amount of heterogeneity increases. A similar pattern holds in comparing the utility of sibling pairs and inbred individuals. Sibling pairs prove to be much more powerful in the presence of a high level of phenocopies, but if phenocopies and susceptibility alleles are rare, inbred individuals are more efficient, with the efficiency increasing as heterogeneity increases.

Our overall conclusion about study design is that sibling pairs are likely to be more useful than either inbred individuals or more distant relative pairs for a fairly wide range of recessive and partially recessive traits, especially if there is only one locus. But in the presence of heterogeneity, inbred individuals can be much more useful. Of course, it may be desirable to use inbred individuals in any case if such data are readily available. If so, our results imply that it is important to define the trait and the population carefully, in order to eliminate phenocopies as much as
possible. In addition, if the inbred individuals come from a special (e.g. inbred) sub-population, care must be taken to evaluate properly the population HBD probability before analyzing the data.

This paper has emphasized the use of our models, particularly the Gaussian models, to obtain qualitative results about mapping strategies. But we have also presented the necessary formulas for analyzing data (computing p-values) from such studies. One practical data analysis issue we have not addressed is that of combining data from different types of inbred individuals and from inbred individuals and siblings. The methods we have described can be extended to a combined dataset without too much difficulty.

If different types of inbred individuals are to be combined (e.g. first and second cousins), the most straightforward approach is simply to look for the maximum of the stochastic process which tracks the total number of individuals who are HBD, regardless of the mating type. The p-value for this statistic will be somewhat more complex than equation (1), but it can be derived similarly using the methods of Feingold (1993). As we have noted, however, different types of inbred individuals have different mapping powers. This suggests that a more optimal approach than just adding the different types together is to use a weighted sum. The problem of finding the best weighting is mathematically similar to the problem of finding the best statistic for sibling pairs. In both cases it is a matter of choosing the appropriate "extreme" of a multi-dimensional process. As with sibling pairs, the optimal weighting depends on the trait etiology, so the best hope is to find a weighting that is fairly efficient for a wide variety of etiologies, as the $T_r(1/4)$ statistic is for sibling pairs. In the context of the Gaussian approximations, it is no
problem to give p-values for statistics with arbitrary weightings, since a weighted average of Gaussian processes is itself a Gaussian process. If sample sizes are small, however, the Markov chain approximations may be more appropriate. Then arbitrary weightings lead to new mathematical problems, which are also encountered in a Markov chain analysis of $T_{\phi}(1/4)$. These problems will be discussed in future work by Tu and Siegmund.

One scenario that is of particular interest for inbred individuals is the case of a rare heterogeneous trait with a negligible level of phenocopies. In that case we have at any particular trait locus $\gamma = \pi$, where $\pi$ is a proportion giving the cases attributable to that particular locus. We can write the log likelihood ratio and examine numerically the relative weights given to the subgroups as a function of $\pi$. For children of first and second cousins, the score statistic, which corresponds to very small $\pi$, would give four times as much weight to the second cousins. However, for values of $\pi$ in the range 0.25 to 0.75, relative weights range from about 4/3 to 5/3. This suggests that a weight in this range might perform quite well for the values of $\pi$ most likely to be interesting. However, some preliminary simulations indicate that little power is lost by using weights of 1/1, i.e., simply adding the first and second cousins together.

The problem of combining data from inbred individuals and sibling pairs is again mathematically similar to choosing the best sibling pair statistic or the best weighting of different types of inbred individuals. If parameter values ($\alpha, \delta$, and all $\gamma$s) are completely specified, the likelihood ratio can be examined to find the optimal weighting of the statistics for each set of data. The practical challenge is to
find weightings (or, alternatively, a non-linear combination of the statistics such as a chi-square type statistic) that are relatively efficient over a wide range of parameter values. Most of the cases where both siblings and inbred individuals are likely to be useful have \( \delta / \alpha \) large, so the statistic \( T_j(0) \), which counts one for each sib pair having identity by descent on both chromosomes, is likely to be the statistic of choice for the sib pairs. This can easily be combined with a homozygosity by descent statistic by simply adding the total number of sib pairs scoring one to the number of inbred individuals scoring one.
Acknowledgements

We wish to thank David Botstein and Michele Gschwend for discussions of homozygosity mapping and in particular of Fanconi anaemia, and Leonid Kruglyak for his comments on a preliminary version of the manuscript.

This work is supported in part by NIH grant R01 - HG00848 - 01A1.
Appendix

Power Approximations for Markov Chain Models

To approximate the power of any of the tests based on Markov models, we write the power as \( P(Y_r \geq b) + P(Y_r < b, Y_t \geq b \text{ for some } t \neq r) \), where \( r \) is the trait locus and \( Y \) is the process of interest. The first term of this expression is generally easy to evaluate. In all cases we have considered, \( Y_r = \sum_{i=1}^{N} Y_r^{(i)} \). For all of the inbred individuals, and for the sibling statistics \( X_{2, t} + X_{1, t} \) and \( X_{2, r} Y_r \), has a binomial distribution. For \( 2X_{2, t} + X_{1, r} \), \( Y_r^{(i)} \) takes just three possible values and is relatively symmetrical, even under the alternative hypothesis. Thus the first term of the power expression can be calculated exactly as a binomial tail probability or, if the sample size is reasonably large, approximately by a normal approximation with continuity correction.

The second term of the power expression can be approximated by

\[
\sum_{i=0}^{k-1} P(Y_r = i) (2V_i - V_i^2),
\]

where \( V_i \) is approximately equal to

\[
V_i = \left( \frac{Q(b-1, b)}{Q(b-1, b-2)} \right)^{b-i},
\]

which comes from assuming that \( Y \) acts like a simple random walk near \( b \). The quantities
$Q(b - 1, b)$ and $Q(b - 1, b - 2)$ are the instantaneous average transition rates away from state $b - 1$, as described in Feingold (1993), with the exception that they must be calculated under the alternative hypothesis. For the statistic $X_{2, \nu}$, 

$$Q(b - 1, b - 2) = 4\lambda(b - 1)$$ and 

$$Q(b - 1, b) = 4\lambda(N - b + 1)[(1 - \delta)(3 - \alpha - 2\delta)].$$

For $X_{2, \nu} + X_{1, \nu}$, 

$$Q(b - 1, b - 2) = 4\lambda(b - 1)[(1 - \delta)(3 + \alpha)]$$ and 

$$Q(b - 1, b) = 4\lambda(N - b + 1).$$ For $2X_{2, \nu} + X_{1, \nu}$, 

$$Q(b - 1, b - 2) = 2\lambda(b - 1)$$ and 

$$Q(b - 1, b) = 2\lambda(2N - b + 1),$$ the same as under the null hypothesis. For inbred individuals, the values of $Q$ are the same under the null and alternative hypotheses in all cases, if we are willing to assume there is only one linked gene per chromosome.

If the sample size is reasonably large, the computation of (A1) can be simplified with a continuous approximation, which is especially useful when $P(Y_r = i)$ is not a binomial probability. The continuous approximation is

$$2\int_{-\infty}^{\infty} \frac{1}{2} \phi\left(\frac{x - \mu}{\tau}\right) V_x dx - \int_{-\infty}^{b - 1/2} \frac{1}{2} \phi\left(\frac{x - \mu}{\tau}\right) V_x^2 dx$$

$$= 2\exp\left[\rho(\mu - b) + \frac{1}{2} \rho^2 \tau^2\right] \times \Phi\left[\frac{b - \frac{1}{2} \mu - \tau^2 \rho}{\tau}\right]$$

$$- \exp[2\rho(\mu - b) + 2\rho^2 \tau^2] \times \Phi\left[\frac{b - \frac{1}{2} \mu - 2\tau^2 \rho}{\tau}\right],$$

where $\mu$ and $\tau$ are the mean and standard deviation of $Y_r$ and
\[ \rho = \ln \left( \frac{Q(b-1,b)}{Q(b-1,b-2)} \right). \]

**Likelihood Ratio Test for Sibling Pairs**

Holmans (1994) discusses the generalized likelihood ratio test, which he calls the "possible triangle test" to reflect the natural genetic constraint \(0 \leq \delta \leq \alpha\).

To describe a Gaussian approximation to the likelihood ratio statistic, it is convenient to introduce the notation

\[ U_{1,t} = -2(X_{1,t} - N/2)/N^{1/2}, \quad U_{2,t} = 2^{1/2}(X_{2,t} - X_{0,t})/N^{1/2}. \]

Direct calculations show that on unlinked chromosomes these two processes have expectation 0, variance 1, and are uncorrelated. Also

\[ \text{Cov}[U_{1,s}, U_{1,t}] = \exp[-\beta_1]t - s, \quad \text{Cov}[U_{2,s}, U_{2,t}] = \exp[-\beta_2]t - s, \]

where \(\beta_1 = 0.08\) and \(\beta_2 = 0.04\). On a linked chromosome having a single trait locus at \(r\),

\[ EU_{1,t} = \mu_1 \exp[-\beta_1]t - s, \quad EU_{2,t} = \mu_2 \exp[-\beta_2]t - s, \quad \text{(A2)} \]

where \(\mu_1 = N^{1/2}\delta\) and \(\mu_2 = (N/2)^{1/2}(\alpha + \delta)\). The genetic constraint \(0 \leq \delta \leq \alpha\) is
equivalent to \(0 \leq 2^{1/2}\mu_1 \leq \mu_2\). Thus the triangular constraint noted by Holmans becomes the constraint that \(\mu = (\mu_1, \mu_2)\) lie in a wedge in the first quadrant of the \(xy\) plane defined by the lines \(y = 2^{1/2}x\) and \(x = 0\). In the case \(\delta = 0\), i.e., an additive model of inheritance, \(U_1\) carries no information, so a test to detect linkage is based directly on the maximum over all loci \(t\) of

\[
U_{2,t}.
\]

(A3)

For a rare recessive trait, where \(\delta = \alpha\), the appropriate test is based on the maximum over \(t\) of

\[
4[X_{2,t} - N/4](3N)^{1/2} = [U_{1,t} + 2^{1/2}U_{2,t}]^{3/2}.
\]

(A4)

The statistic (A4) is the projection of the vector \((U_{1,t}, U_{2,t})\) along the line \(y = 2^{1/2}x\), which makes an angle \(\tan^{-1}2^{1/2}\) with the positive \(x\) axis in the \(xy\) plane. The statistic (A3) is obviously the projection of \((U_{1,t}, U_{2,t})\) along the \(y\) axis, which makes an angle of \(\pi/2\) with the positive \(x\) axis.

The log likelihood function is

\[
\mu_1 U_{1,t} + \mu_2 U_{2,t} - (\mu_1^2 + \mu_2^2)/2.
\]

If there were no constraints on \((\mu_1, \mu_2)\), the likelihood ratio statistic at the locus \(t\) would be obtained by maximizing this expression with respect to \((\mu_1, \mu_2)\), which yields
\[ [U_{1,t}^2 + U_{2,t}^2]^{1/2}. \]  

(A5)

(The square root is introduced to keep (A5) in the same units as (A3) and (A4).) To incorporate the constraints we use (A5) if the point \((U_{1,t}, U_{2,t})\) lies in the wedge defined by the lines \(y = 2^{1/2}x\) and \(x = 0\). If the point does not lie in this wedge, we use the larger of (A3) and (A4). In effect the likelihood ratio test incorporating the constraints is based on (A5) unless the data tell us that the mode of inheritance appears to be purely additive or purely recessive; in these extreme cases we use the statistic appropriate for the apparent mode of inheritance. The false positive rate of the likelihood ratio test is the probability, computed under the assumption \(\alpha = \delta = 0\) that the statistic described above exceeds the detection threshold \(b\) somewhere throughout the genome. This probability can be evaluated approximately by an argument along the lines given in Feingold et al. (1993), although more complicated. Details of a related calculation are given by Dupuis (1994). For a genome of length \(l\) the approximation is

\[ 1 - \exp\left[-le^{b^2/2}\left(Cb^2/(2\pi) + b(5\beta_2/6 + \beta_1/6)(2\pi)^{1/2}\right)\right], \]

where

\[ C = \int_{\pi^{1/2}}\int_{-1}^{1} (\beta_2 \sin^2 \omega + \beta_1 \cos^2 \omega) d\omega. \]
This integral is easily shown to equal

\[
(b_1 + b_2)(\pi/2 - \tan^{-1}2^{1/2})/2 - (b_1 - b_2)(3 \times 2^{1/2}) = 0.0275 \text{ for } b_1 = 0.08 \text{ and } b_2 = 0.04. \text{ For a genome of genetic length } 3000 \text{ cM a threshold of } b = 4.3 \text{ yields a false positive error rate of approximately } 0.05.
\]

To approximate the power of the likelihood ratio test we let \( \xi = [\mu_1^2 + \mu_2^2]^{1/2} \) denote the distance of the point \( \mu = (\mu_1, \mu_2) \) from the origin. The point \( \mu \) lies inside the wedge defined by the angles \( \tan^{-1}2^{1/2} \) and \( \pi/2 \) measured from the positive \( x \) axis; and a slightly different approximation to the power is appropriate depending on whether the point is strictly inside the wedge or on one of the edges. In the former case the power is approximately

\[
1 - \Phi(b - \xi) + \phi(b - \xi) \left(\frac{1}{2\xi} + \frac{\xi}{2\xi^2 - 1/(b + \xi)}\right) \tag{A6}
\]

while in the latter it is

\[
1 - \Phi(b - \xi) + \phi(b - \xi) \left(\frac{1}{4\xi} + \frac{1}{1/\xi - 1/(2(b + \xi))}\right) \tag{A7}
\]

These approximations can be derived by arguments similar to those used in Feingold et al. (1993) and Siegmund (1985). Details will be given elsewhere.

The likelihood ratio test behaves much like the test based on \( T_\alpha(1/4) \) discussed above. That test was chosen to be optimal in the case \( \delta = \alpha/3 \) or equivalently \( V_D = V_A \), which is intermediate between an additive and a recessive mode of inheritance. As a result it is slightly more efficient than the likelihood ratio
test when the mode of inheritance is indeed intermediate, and slightly less efficient when the mode of inheritance is closer to one of the two extremes. The minimum efficiency for both tests occurs at the extreme models, where the likelihood ratio test performs slightly better. For example, to achieve power at least 0.9 regardless of the mode of inheritance, the optimal test if one knew the value of $\delta/\alpha$ would require a value of $\xi = 5.0$. The likelihood ratio test would require $\xi = 5.2$, and the test based on $T_{A}(1/4)$ would require $\xi = 5.25$. Since the parameter $\xi$ is proportional to the square root of the sample size the relative efficiencies of these tests are proportional to the squares of these parameter values. In particular the likelihood ratio test is about 92% efficient, while the test based on $T_{A}(1/4)$ is about 91% efficient compared to the optimal test based on the true value of $\delta/\alpha$.

**Maximum Test for Sibling Pairs**

Schaid and Nick (1990) suggest using as a statistic the maximum of (A3) and (A4). Some simple calculations not given here suggest that the performance of this test is similar to that of the likelihood ratio test and the test based on $T_{A}(1/4)$. To the extent that it differs from these other two, it will be comparatively better if one of the extreme models $\delta = 0$ or $\delta = \alpha$ holds and comparatively weaker for intermediate modes of inheritance.
References

Bodmer WF, Cavalli-Sforza LL (1976) Genetics, evolution and man. W H Freeman, San Francisco


Unpublished manuscript.


Knapp, Seuchter, Baur (1994b) Linkage analysis in nuclear families. 2: Relationship between affected sib-pair tests and lod score analysis. Hum Hered 44:44-51


Lander ES, Botstein D (1986) Strategies for studying heterogeneous genetic traits
in humans by using a linkage map of restriction fragment length polymorphisms. Proc Nat Acad Sci USA 83: 7353-7357


Risch N (1990b) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229-241

Schaid DJ, Nick TG (1990) Sib-pair linkage tests for disease susceptibility loci: common tests vs. the asymptotically most powerful test. Genet Epi 7:359-370

Siegmund D (1985) Sequential analysis: tests and confidence intervals. Springer-Verlag, New York


### Table 1  P-value parameters for different mating types.

<table>
<thead>
<tr>
<th>inbreeding type</th>
<th>affected-pair type</th>
<th>$p_0 (= F)$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sibling</td>
<td>cousin</td>
<td>1/4</td>
<td>$16\lambda/3$</td>
</tr>
<tr>
<td>avuncular</td>
<td>cousin once removed</td>
<td>1/8</td>
<td>$40\lambda/7$</td>
</tr>
<tr>
<td>cousin</td>
<td>second cousin</td>
<td>1/16</td>
<td>$32\lambda/5$</td>
</tr>
<tr>
<td>cousin once removed</td>
<td>second cousin once removed</td>
<td>1/32</td>
<td>$224\lambda/31$</td>
</tr>
<tr>
<td>second cousin</td>
<td>third cousin</td>
<td>1/64</td>
<td>$512\lambda/63$</td>
</tr>
</tbody>
</table>

### Table 2  Values of $R(x)$ for different mating types.

<table>
<thead>
<tr>
<th>inbreeding type</th>
<th>$R(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>parent/child</td>
<td>$\frac{1}{3}e^{-2\lambda x} + \frac{1}{3}e^{-4\lambda x} + \frac{1}{3}e^{-6\lambda x}$</td>
</tr>
<tr>
<td>half-sibling</td>
<td>$\frac{2}{3}e^{-2\lambda x} + \frac{1}{3}e^{-4\lambda x} + \frac{1}{3}e^{-6\lambda x} + \frac{1}{3}e^{-8\lambda x}$</td>
</tr>
<tr>
<td>sibling</td>
<td>$\frac{3}{6}e^{-4\lambda x} + \frac{1}{6}e^{-6\lambda x} + \frac{1}{6}e^{-8\lambda x}$</td>
</tr>
<tr>
<td>uncle/niece</td>
<td>$\frac{2}{14}e^{-2\lambda x} + \frac{3}{14}e^{-4\lambda x} + \frac{5}{14}e^{-6\lambda x} + \frac{3}{14}e^{-8\lambda x} + \frac{1}{14}e^{-10\lambda x}$</td>
</tr>
<tr>
<td>cousin</td>
<td>$\frac{4}{30}e^{-2\lambda x} + \frac{6}{30}e^{-4\lambda x} + \frac{8}{30}e^{-6\lambda x} + \frac{6}{30}e^{-8\lambda x} + \frac{4}{30}e^{-10\lambda x} + \frac{1}{30}e^{-12\lambda x}$</td>
</tr>
</tbody>
</table>
Table 3  Example parameter combinations for a two-allele model. $f_0 + 2f$ is the penetrance of the homozygote. $K$ is the trait frequency. $f_0/K$ is the fraction of cases due to phenocopies. $-dlf$ is the degree of recessivity under the two-allele model. $\gamma_1$ and $\gamma_2$ are the parameters for offspring of first and second cousins, respectively.

<table>
<thead>
<tr>
<th># loci</th>
<th>$f_0 + 2f$</th>
<th>$-dlf$</th>
<th>$K$</th>
<th>$f_0/K$</th>
<th>$\alpha$</th>
<th>$\delta$</th>
<th>$\gamma_1$</th>
<th>$\gamma_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.00002</td>
<td>0</td>
<td>1.0</td>
<td>.98</td>
<td>.93</td>
<td>.78</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.0002</td>
<td>0</td>
<td>1.0</td>
<td>.95</td>
<td>.81</td>
<td>.52</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.0002</td>
<td>.5</td>
<td>1.0</td>
<td>.96</td>
<td>.77</td>
<td>.43</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.002</td>
<td>0</td>
<td>.99</td>
<td>.84</td>
<td>.57</td>
<td>.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.002</td>
<td>.5</td>
<td>.99</td>
<td>.87</td>
<td>.49</td>
<td>.19</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.01</td>
<td>0</td>
<td>.97</td>
<td>.67</td>
<td>.36</td>
<td>.12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.01</td>
<td>.5</td>
<td>.93</td>
<td>.72</td>
<td>.29</td>
<td>.09</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.02</td>
<td>0</td>
<td>.94</td>
<td>.57</td>
<td>.28</td>
<td>.09</td>
</tr>
<tr>
<td>1</td>
<td>.5</td>
<td>1</td>
<td>.0002</td>
<td>0</td>
<td>1.0</td>
<td>.92</td>
<td>.75</td>
<td>.43</td>
</tr>
<tr>
<td>1</td>
<td>.5</td>
<td>1</td>
<td>.01</td>
<td>0</td>
<td>.94</td>
<td>.56</td>
<td>.28</td>
<td>.09</td>
</tr>
</tbody>
</table>

Table continued on next page
<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>0.9</th>
<th>0.0002</th>
<th>0</th>
<th>0.99</th>
<th>0.13</th>
<th>0.36</th>
<th>0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.9</td>
<td>0.01</td>
<td>0</td>
<td>0.93</td>
<td>0.48</td>
<td>0.25</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.0002</td>
<td>0</td>
<td>0.50</td>
<td>0.49</td>
<td>0.48</td>
<td>0.42</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.0002</td>
<td>0</td>
<td>0.50</td>
<td>0.48</td>
<td>0.43</td>
<td>0.31</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.002</td>
<td>0.5</td>
<td>0.49</td>
<td>0.45</td>
<td>0.30</td>
<td>0.14</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.01</td>
<td>0</td>
<td>0.48</td>
<td>0.37</td>
<td>0.23</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.02</td>
<td>0</td>
<td>0.47</td>
<td>0.32</td>
<td>0.18</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.02</td>
<td>0.5</td>
<td>0.44</td>
<td>0.34</td>
<td>0.15</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0.0002</td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0.0002</td>
<td>0</td>
<td>0.25</td>
<td>0.24</td>
<td>0.23</td>
<td>0.18</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0.01</td>
<td>0</td>
<td>0.24</td>
<td>0.20</td>
<td>0.13</td>
<td>0.06</td>
</tr>
</tbody>
</table>
Table 4  Sample sizes for selected parameter values. All loci contribute equally. Power to detect any particular locus is 90%.

<table>
<thead>
<tr>
<th># loci</th>
<th>$\alpha$</th>
<th>$\delta$</th>
<th>sib pairs</th>
<th>1st cousins</th>
<th>2nd cousins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.95</td>
<td>9</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>.99</td>
<td>.84</td>
<td>12</td>
<td>15</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>.97</td>
<td>.67</td>
<td>16</td>
<td>29</td>
<td>92</td>
</tr>
<tr>
<td>1</td>
<td>.93</td>
<td>.72</td>
<td>16</td>
<td>42</td>
<td>140</td>
</tr>
<tr>
<td>1</td>
<td>.93</td>
<td>.48</td>
<td>25</td>
<td>55</td>
<td>168</td>
</tr>
<tr>
<td>2</td>
<td>.5</td>
<td>.49</td>
<td>42</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>.5</td>
<td>.48</td>
<td>44</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>.5</td>
<td>.44</td>
<td>49</td>
<td>35</td>
<td>58</td>
</tr>
<tr>
<td>2</td>
<td>.49</td>
<td>.45</td>
<td>49</td>
<td>41</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>.48</td>
<td>.37</td>
<td>63</td>
<td>63</td>
<td>169</td>
</tr>
<tr>
<td>2</td>
<td>.47</td>
<td>.32</td>
<td>74</td>
<td>95</td>
<td>257</td>
</tr>
<tr>
<td>2</td>
<td>.44</td>
<td>.34</td>
<td>74</td>
<td>124</td>
<td>330</td>
</tr>
<tr>
<td>4</td>
<td>.25</td>
<td>.25</td>
<td>148</td>
<td>55</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>.25</td>
<td>.24</td>
<td>164</td>
<td>63</td>
<td>61</td>
</tr>
<tr>
<td>4</td>
<td>.24</td>
<td>.20</td>
<td>211</td>
<td>161</td>
<td>257</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1  Performance of the statistic $T(c)$ for $c = 1/2, 1/4,$ and 0, expressed in terms of the percentage increase needed in the sample size compared to the ideal value of $c$.

Figure 2  Relative efficiency of sibling pairs to cousin pairs. Values for $\lambda_S = \lambda_O$ are computed using the sibling statistic $T(1/2)$. Values for $\lambda_O = (1 + \lambda_S)/2$ are computed using $T(1/4)$.

Figure 3  Schematic of pattern of HBD in offspring of a parent/child mating.
Figure 1

% above sample size for ideal statistic

- $c = 0$
- $c = 1/4$
- $c = 1/2$

$k$
Figure 2

$\lambda_O = (1 + \lambda_S)/2$

$\lambda_O = \lambda_S$
Figure 3

chromosome 1
chromosome 2