Am. J. Hum. Genet. 58:844–860, 1996

A General Statistical Model for Detecting Complex-Trait Loci by Using Affected Relative Pairs in a Genome Search

Susan L. Smalley, J. Arthur Woodward, and Christina G. S. Palmer*

University of California, Los Angeles

Summary

Scanning of the human genome by use of affected relative pairs and dense sets of highly polymorphic markers or by emerging techniques such as genomic mismatch scanning (GMS) is making it possible to identify the genetic etiology of a disease through detection of susceptibility loci. We present a general statistical model and test to detect disease genes, using affected relative pairs and either markers or GMS technologies in a genome search. There are an exact test and large-sample normal approximation that control for the elevated probability of false detection of linkage in a genome search. The approach can be used to determine the sample size needed to obtain a prespecified power to detect a disease gene in the presence of etiologic heterogeneity for a single class or mixture of relative classes, with any number of markers or clones, marker PIC values, or mapping function. The approach is used to examine differences in performance of markers and GMS technologies in a common statistical framework and to provide practical information for designing studies of complex traits.

Introduction

Conducting a genome search by use of affected-relative-pair methods is emerging as a major tool for the genetic dissection of a complex trait (Lander and Schork 1994). Success of the approach depends, in part, on the availability of highly polymorphic, densely spaced markers and on efficient methods of genotyping these markers. Advances in molecular technology, including development of novel methodologies, such as genomic mismatch scanning (GMS) (Nelson et al. 1993), are rapidly reducing genotyping costs and increasing efficiency, making feasible the use of a genome search as an exploratory tool in identifying the genetic etiology of disease. The use of a genome search and affected-pair sampling as an exploratory tool may require evaluation of hundreds of polymorphic markers or, in the case of GMS, thousands of clones, which, in the absence of susceptibility loci, can lead to an elevated probability of false detection of linkage.

In general, linkage methods are applied once the genetic etiology of a trait is well established, and, in the context of a genome search, correct identification of a marker linked to a putative disease gene increases in probability as more and more markers are genotyped (Risch 1991). However, use of a genome search to establish the genetic etiology of a trait by directly localizing the genes differs from the former case, because less a priori information about the genetic etiology of the trait may be available. In fact, Kruglyak and Lander (1995) report that use of a minimum LOD score of 2.3 in a genome search using affected sibling pairs (when there are no susceptibility genes) will yield a type I error with ~.8 probability. As a genome search using affected relative pairs is applied in an exploratory fashion, a statistical approach that handles realistically complex etiologies and that can be used with any sampling design is desirable. Furthermore, it is desirable that the approach be applicable to different molecular technologies currently available or in development, so that their performance in a genome search may be compared in a common statistical framework with explicit control for the type I error.

Currently, a genome search using affected relative pairs requires evaluation of polymorphic markers spanning the genome at fairly regular intervals, to detect increased allele sharing among pairs. If a marker is completely polymorphic—i.e., each chromosome has a unique allele at the marker—increased allele sharing for a particular marker across pairs of affected relatives (above the expected background rate) reflects increased identity by descent (IBD) and the presence of a susceptibility gene underlying that trait (Suarez 1978; Risch 1990*b*). However, for markers that are less than completely polymorphic, increased allele sharing may occur because of identity by state (IBS), as well as because of IBD, so that increased IBD must be inferred (Lange 1986; Bishop and Williamson 1990; Risch 1990*c*).

GMS is a novel molecular technique that circumvents

Received August 11, 1995; accepted for publication January 19, 1996.

Address for correspondence and reprints: Dr. Susan L. Smalley, Department of Psychiatry, 47-438 NPI, UCLA School of Medicine, 760 Westwood Plaza, Los Angeles, CA 90024. E-mail: ssmalley@npimain.medsch.ucla.edu

*The authors contributed equally to this work.

0002-9297/96/5804-0023$02.00

the problem of allele sharing due to IBS, because it enables all regions of IBD in the genome to be identified between affected pairs of relatives (Nelson et al. 1993). The technique is based on the unique ability of DNA from IBD regions to form large "heterohybrid" duplexes (one strand from each individual) that are free of mismatches. Mismatch-free heterohybrids, ~10–20 kb in size, are then hybridized onto either metaphase chromosomes or clones containing contiguous human genomic DNA, to identify chromosome locations of increased IBD across pairs of affected relatives. In a clone-based GMS procedure, thousands of clones are evaluated (e.g., 3,300 clones, each 1 Mb in size, in a genome 3,300 cM in length). Thus, GMS is a technology that is equivalent to having 100% polymorphic markers that completely cover the genome. Although GMS is not yet feasible in eukaryotic organisms, work is in progress at several laboratories to apply it to the human genome.

Statistical methods using affected relative pairs (or pedigrees) and markers in the context of a genome search have been the focus of recent research (Elston 1992; Brown et al. 1994; Kruglyak and Lander 1995). Lander and Schork (1994) provide an excellent review of recent studies addressing difficult issues involved in a genome search, including choice of significance levels, sampling designs, and marker densities.

Statistical methods using GMS technology in the context of a genome search are in the early stages of development. Feingold (1993) and Feingold et al. (1993) have described a set of stochastic processes including Markov chain and Gaussian approximations for analyzing qualitative traits, with pairs of affected relatives. Feingold et al. (1993) used the Orstein-Uhlenbeck process and an assumption of normality to model the dependencies among segments of DNA along a single chromosome, in order to approximate their test, which uses the largest order statistic. Their statistical test is adjusted for the number of independent chromosomes evaluated, by use of the Bonferroni procedure. N. J. Schork and S. Ghosh (personal communication) proposed modification of GMS (MGMS), in order to fully differentiate states of parental allele sharing in siblings. Two MGMS assays were described. One assay applied representational difference analysis on GMS-selected DNA. The other assay compared material selected by GMS with that not selected by GMS, to distinguish between sibling sharing of one parental DNA segment and sibling sharing of two parental DNA segments, at a specified region. They presented a statistical approach to detect quantitative-trait loci by using an MGMS assay and nonoverlapping clones.

We present a general statistical model for a genome search using affected relative pairs and either a marker- or clone-based GMS technology. The approach can be used for mixtures of relative-pair classes in the presence of complex etiologies. There are an exact test and large-sample normal approximation that control for the elevated probability of false detection of linkage in a genome search. By creation of sets that contain independent markers or clones, the exact distribution of the largest order statistic, without a normality assumption, is used to identify significantly increased allele sharing under markers—or enriched regions of IBD under GMS—among independent pairs of affected relatives. The statistical test is adjusted for the number of dependent sets of markers evaluated, by use of the Bonferroni procedure. We illustrate power to detect susceptibility genes for markers and GMS under a variety of genetic models and sampling designs, in order to compare these two technological approaches and to provide practical information for designing a genome search.

Method

The test presented in this paper takes advantage of the fact that we can identify statistically independent markers or clones and group them together into mutually exclusive sets for subsequent analysis. Statistical independence is approximated by including in a set only those markers or clones that are known to assort independently, either because they lie on different chromosomes or because they are $\geq$100 cM (or, for clones, 100 Mb) from each other. Since, under independent assortment, the recombination frequency, θ, is .5, two markers separated by $\geq$100-cM distances are approximately statistically independent. The value of θ corresponding to 100 cM depends on the choice of mapping function (Ott 1991). The fact that markers or clones across sets are statistically dependent does not violate the assumptions of our statistical test.

To illustrate aspects of the approach throughout this paper, we describe two hypothetical genome-search strategies, one for markers and one for clone-based GMS, assuming a 3,300-cM autosomal, sex-averaged, haploid genome (Renwick 1969). For a marker search, we evaluate 330 markers with 10-cM spacing between adjacent markers, such that any disease locus is, at most, ~5 cM from a marker. The 330 markers are grouped into 10 sets of 33 statistically independent markers. For clone-based GMS, we evaluate 3,300 clones, each 1 Mb in size. These 3,300 clones are placed into 100 sets of 33 statistically independent clones. Each set of 33 markers or clones is examined for similarity among n independent pairs of affected relatives.

As originally described, GMS provides complete coverage of the genome (e.g., 3,300 clones); however, a partial GMS (PGMS) procedure could also be implemented, by hybridization to a subset of clones rather than to the entire genome. Such a procedure is equivalent to evaluating completely polymorphic markers at intervals throughout the genome. We will refer to PGMS and completely polymorphic markers covering a frac-

tion of the genome as "interval IBD" (I-IBD) and will consider the genome-search strategy to be that described above for 330 markers.

Measures of Allele Sharing

Let X_{ij} be a random variable whose values represent either hybridization of GMS-selected DNA to clones or sharing of alleles for markers for relative class i ($i = 1, c$) and relative pair j ($j = 1, n_i$), where c is the number of relative classes (e.g., siblings, first cousins, etc.) and n_i is the number of pairs of relatives in class i. The total number of relative pairs is $n = \Sigma_{i=1}^{c} n_i$. In the case of GMS, X_{ij} has only two possible values, 0 (no hybridization) or 1 (hybridization), for all relative classes. Note that, for the sibling case, X_{ij} takes on the value of 1 when either one or two parental alleles are shared IBD, because these two IBD states cannot be differentiated by the GMS technology described by Nelson et al. (1993). For convenience, in the remainder of this paper we use the term "allele sharing" with the understanding that, in the case of a GMS technology, this term refers to hybridization of GMS-selected DNA to clones. In the case of markers and MGMS, X_{ij} has three possible values—0, 1, and 2—which reflect sharing of no alleles, sharing of one allele, and sharing of two alleles, respectively.

The sum of the X_{ij} within the ith class is denoted as $Y_i = \Sigma_{j=1}^{n_i} X_{ij}$, and the sum across all relative classes is denoted as $Z = \Sigma_{i=1}^{c} Y_i = \Sigma_{i=1}^{c} (\Sigma_{j=1}^{n_i} X_{ij})$. When X_{ij} takes on two values, the sum Y_i is equivalent to a statistic described by Blackwelder and Elston (1985), Bishop and Williamson (1990), and Risch (1990*b*). When X_{ij} takes on three states, the sum Y_i is equivalent to the mean statistic described by Green and Woodrow (1977), Blackwelder and Elston (1985), and Thomson and Motro (1994).

The random variable X_{ij} is evaluated for the rth marker or clone ($r = 1, g$) in a set, where g is the total number of markers or clones evaluated in a set and s is the number of sets. Thus, $(g)(s)$ is the total number of markers or clones evaluated in any study. For the marker search strategy outlined above, $g = 33$ and $s = 10$; for the GMS search strategy, $g = 33$ and $s = 100$.

Null distribution of X_{ij}.—In the absence of a disease locus in the genome and, therefore, in any set—the null distribution of X_{ij} at the rth marker or clone is

$$\begin{bmatrix} \pi_{i0} \\ \pi_{i1} \\ \pi_{i2} \end{bmatrix} = \begin{bmatrix} T_{00} & 0 & 0 \\ T_{10} & T_{11} & 0 \\ T_{20} & T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} \kappa_{i0} \\ \kappa_{i1} \\ \kappa_{i2} \end{bmatrix} \qquad (1)$$

$$\boldsymbol{\pi}_i = \mathbf{T}\,\boldsymbol{\kappa}_i ,$$

for family class i, where $\pi_{iu} = P(\text{IBS} = u)$ at marker locus r and T is a lower-triangular matrix with $T_{uv} = P(\text{IBS} = u | \text{IBD} = v)$ for the rth locus. The elements of T are derived from known population allele frequencies at the rth marker locus (Bishop and Williamson 1990, p. 255). Also, in equation (1) $\boldsymbol{\kappa}_i$ is a vector of coefficients of relationship, where, for all r, $\kappa_{iv} = P(\text{IBD} = v)$ for relative class i (Wright 1922). Note that, for unilineal relatives, $\kappa_{i2} = 0$. For GMS, MGMS, and I-IBD, T = I, the identity matrix, because at marker locus r the $P(\text{IBS} = u | \text{IBD} = v)$ equals 1 when $u = v$ and equals 0 when $u \neq v$.

Null distribution of Y_i.—In the case of GMS, Y_i is binomial and denoted

$$Y_i \sim \text{Bin}(n_i, \pi_{i1} + \pi_{i2}) . \qquad (2)$$

In large samples, the null distribution of Y_i can be approximated by the normal distribution with expected value and variance:

$$\begin{aligned} \text{E}(Y_i | H_0) &= n_i(\pi_{i1} + \pi_{i2}) \\ \text{Var}(Y_i | H_0) &= n_i(\pi_{i1} + \pi_{i2})[1 - (\pi_{i1} + \pi_{i2})] . \end{aligned} \qquad (3)$$

For markers and MGMS, the distribution of Y_i is no longer binomial, because all three states of allele sharing are present. We refer to the distribution of the sum of n_i independent three-level discrete random variables as

$$Y_i \sim S(n_i, \pi_{i1}, \pi_{i2}) . \qquad (4)$$

We retain the use of all three states of allele sharing for siblings, when using MGMS or markers, because this measure of allele sharing has been shown generally to have equal or greater power than one that dichotomizes X_{ij} (Blackwelder and Elston 1985). Our own comparisons of alternate measures showed a similar finding for the range of disease-allele frequencies considered in this paper. The exact distribution of Y_i is computed by use of the computational algorithm given in appendix A. In large samples, Y_i can be approximated by the normal distribution with expected value and variance

$$\begin{aligned} \text{E}(Y_i | H_0) &= n_i(\pi_{i1} + 2\pi_{i2}) \\ \text{Var}(Y_i | H_0) &= n_i[\pi_{i1}(1 - \pi_{i1}) \\ &\quad + 4\pi_{i2}(1 - \pi_{i2}) - 4\pi_{i1}\pi_{i2}] . \end{aligned} \qquad (5)$$

For unilineal relatives we combine the allele-sharing states of 1 and 2 into a single category, because our previous power comparisons revealed that this measure of allele sharing has similar or greater power than one using all three states. Hence, for all power computations involving unilineal relatives, we consider the distribution of Y_i to be binomial—i.e., $\text{Bin}(n_i, \pi_{i1} + \pi_{i2})$, as defined in equation (2) for the case of GMS. The expected value

and variance for the normal approximation are found in equation (3).

Null distribution of Z.—The sum of Y_i across all relative classes, $Z = Y_1 + Y_2 + \ldots + Y_c$, is the convolution of discrete random variables, because each relative class has unique probabilities of allele sharing (i.e., π_i). The density function of a sum of independent random variables, i.e., $f(Z)$, can be expressed as a function of the densities of the components, i.e., $f(Y_i)$; this is called the "convolution of densities" (Feller 1957) and is denoted as

$$f(Z) = f(Y_1)*f(Y_2)* \ldots *f(Y_c) . \quad (6)$$

For GMS, the null distribution of Z is the convolution of the c binomials given in equation (2). For a study using MGMS for siblings and GMS for the $(c - 1)$ unilineal relative classes, the null distribution of Z is the convolution of the single distribution of Y_i in equation (4) and the $(c - 1)$ binomial distributions in equation (2); the same convolution would apply for a study using markers and the same mixture of relative classes. The exact convolution of densities in equation (6) can be computed efficiently by use of the convolution algorithm in appendix A. In large samples the density of Z can be approximated by the normal with $E(Z) = \Sigma_{i=1}^{c} E(Y_i|H_0)$ and $Var(Z) = \Sigma_{i=1}^{c} Var(Y_i|H_0)$, by use of the expected values and variances of Y_i found in equations (3) and (5).

Statistical Test

The test statistic is $Z' = \max(Z_1, Z_2, \ldots, Z_g)$, the largest order statistic in a set of g independent markers or clones. Our use of the largest order statistic follows Feingold et al. (1993). The statistical test for detecting increased allele sharing is

$$\begin{aligned} &\text{accept } H_0 \text{ if } z' < b , \\ &\text{accept } H_1 \text{ if } z' \geq b , \end{aligned} \quad (7)$$

where b is the value exceeded by the α_{pc} proportion of the null distribution of Z', and α_{pc} is the per-comparison type I error probability. The desired experiment-wise type I error probability, α_{ew}, and the total number of sets evaluated in any study are used to determine the α_{pc}, where $\alpha_{pc} = \alpha_{ew}/s$ (Hochberg and Tamhane 1987). For example, with 330 markers arranged into 10 sets of 33 markers, the Z' in each set are evaluated at $\alpha_{pc} = .05/10 = .005$. For a GMS technology, with 3,300 clones arranged into 100 sets of 33 clones each, the Z' in each set are evaluated at $\alpha_{pc} = .05/100 = .0005$.

Under the null hypothesis of no susceptibility locus, the distribution of the largest order statistic is

$$F(Z':b) = P(Z' < b) = F(Z:b)^g , \quad (8)$$

under the assumption that there is independence of the g markers or clones within a set. The critical value of the test in equation (7) is computed by inverting the cumulative density function in equation (8) by use of a discrete bisection algorithm (e.g., see Knuth 1973), in order to find the b value that yields a prespecified value of the cdf. We note that use of the notation $F(Z:b)^g$ implies that the distribution of each Z is identical, which will be true when $\mathbf{T} = \mathbf{I}$ (e.g., under GMS). In the case of markers, when $\mathbf{T} \neq \mathbf{I}$, the distribution of Y_i at each marker g is different, because of the marker-specific allele number and frequency. To adjust for different marker polymorphisms, each marker-specific Y_i is expressed in standard form, in terms of its own mean and variance. Thus, for each of the g markers, the convolution, Z, of the class-specific standardized Y_i's have identical null distributions.

Alternative Distributions and Power

We now define the alternative distributions of the measure of allele sharing for a pair (X_{ij}), the sum within the ith relative class (Y_i), and the sum across relative classes (Z), when at least one gene underlies the disease trait.

Alternative distribution of X_{ij}.—In the presence of disease gene k, the alternative distribution of X_{ij} for a marker or clone at or near the gene is

$$\begin{bmatrix} \Psi_{i0} \\ \Psi_{i1} \\ \Psi_{i2} \end{bmatrix} = \begin{bmatrix} T_{00} & 0 & 0 \\ T_{10} & T_{11} & 0 \\ T_{20} & T_{21} & T_{22} \end{bmatrix} \times \begin{bmatrix} D_{i00} & D_{i01} & D_{i02} \\ D_{i10} & D_{i11} & D_{i12} \\ D_{i20} & D_{i21} & D_{i22} \end{bmatrix} \begin{bmatrix} \lambda_{i0} \\ \lambda_{i1} \\ \lambda_{i2} \end{bmatrix} , \quad (9)$$

$$\boldsymbol{\Psi}_i = \mathbf{T}\mathbf{D}_i \boldsymbol{\lambda}_i$$

where, for family class i, $\Psi_{iu} = P(\text{IBS} = u|2 \text{ affected relatives}, \theta)$. Thus, the matrix $\boldsymbol{\Psi}_i$ contains the conditional IBS probabilities for marker locus r linked to disease locus k, given two affected relatives and recombination frequency θ. The $P(2 \text{ affected relatives})$, denoted "$P(2$ AR)," is a function of disease-allele frequency, within- and between-locus interaction, and degree of genetic relationship; definition of this probability is found, in the notation used in this paper, in appendix B. The T matrix, previously defined in equation (1), is the same under the null and alternative distributions of X_{ij}, because the elements are a function of marker polymorphism only. The $\mathbf{D}_i$ matrix contains conditional IBD probabilities for a disease locus k, given marker locus r for the ith relative class, under the assumption of linkage equilibrium (Bishop and Williamson 1990, p. 258; Risch 1990*b*, p. 231). The vector $\boldsymbol{\lambda}_i$ contains conditional IBD probabilities for the ith relative class

with element $\lambda_{iu} = P(\text{IBD} = u|2\text{ AR})$ and $\lambda_{i2} = 0$ for unilineal relatives. Definition of $\boldsymbol{\lambda}_i$, T, and $\mathbf{D}_i$ can be found in other sources (e.g., Risch 1990*b*) and appear, in terms of the notation of this paper, in appendix B.

Since the vector $\boldsymbol{\Psi}_i$ is, in part, a function of θ (via $\mathbf{D}_i$) and the PIC of the markers (T), it is immediately possible to illustrate differences among genome-search technologies. For example, with use of GMS, the alternative distribution of X_{ij} is obtained from equation (9) by setting $\mathbf{D}_i = \mathbf{T} = \mathbf{I}$, in which case, $\Psi_{iu} = \lambda_{iu} = P(\text{IBD} = u|2\text{ AR})$ at locus k. In contrast, with use of I-IBD, $\mathbf{T} = \mathbf{I}$ but $\mathbf{D}_i \neq \mathbf{I}$; and, with use of markers, $\mathbf{T} \neq \mathbf{I}$ and $\mathbf{D}_i \neq \mathbf{I}$.

Alternative distribution of Y_i.—The alternative distribution of Y_i, the sum of the X_{ij} within relative class i, depends on the complexity of the disease trait under study. In the case of a single gene with no phenocopies, Y_i is distributed either as a single binomial, $\text{Bin}(n_i, \Psi_{i1}+\Psi_{i2})$, or as the sum of three-level discrete random variables, $S(n_i, \Psi_{i1}, \Psi_{i2})$.

In the presence of etiologic heterogeneity, including phenocopies and l multiple loci, the sample of affected pairs within a relative class is actually composed of two different groups. The first group is composed of those pairs in which both members carry the disease genotype at locus k and whose probabilities of allele sharing are from the alternative distribution of X_{ij} in equation (9). The second group contains those pairs in which at least one member of the pair does not have locus k and whose probabilities of allele sharing at locus k are from the null distribution of X_{ij} in equation (1). The proportion of the sample in which both members carry the disease gene at locus k is denoted as β_i, whereas $(1 - \beta_i)$ is the proportion of the sample in which at least one member of a pair either is a phenocopy or is affected because of a different disease locus, in the case of $l > 1$. The proportion, β_i, is a function of the disease-allele frequency, the mechanism of gene action, between-locus interaction, and degree of genetic relationship. Definition of β_i under epistasis, locus heterogeneity, and in the presence of phenocopies is in appendix B.

In the presence of etiologic heterogeneity ($\beta_i < 1$), the alternative distribution of Y_i at or near the kth disease gene is itself a convolution of two independent binomials or three-level discrete random variables, depending on the number of states of X_{ij}. When X_{ij} takes on two values (i.e., GMS or unilineal relatives with markers), the distribution of Y_i is the convolution of two independent binomials,

$$Y_i \sim \text{Bin}(n_i\beta_i, \Psi_{i1} + \Psi_{i2}) * \text{Bin}[n_i(1 - \beta_i), \pi_{i1} + \pi_{i2}] . \quad (10)$$

In this equation and similar ones that follow, $n_i\beta_i$ is rounded to the closest integer. In large samples, the distribution of Y_i can be approximated by the normal with expected value and variance,

$$E(Y_i|H_1) = n_i\beta_i(\Psi_{i1} + \Psi_{i2}) + n_i(1 - \beta_i)(\pi_{i1} + \pi_{i2})$$

$$\text{Var}(Y_i|H_1) = n_i\beta_i(\Psi_{i1} + \Psi_{i2})[1 - (\Psi_{i1} + \Psi_{i2})] + n_i(1 - \beta_i)(\pi_{i1} + \pi_{i2})[1 - (\pi_{i1} + \pi_{i2})] . \quad (11)$$

When X_{ij} takes on three values (i.e., MGMS or markers with siblings), the distribution of Y_i is the convolution

$$Y_i \sim S(n_i\beta_i, \Psi_{i1}, \Psi_{i2}) * S[n_i(1 - \beta_i), \pi_{i1}, \pi_{i2}] . \quad (12)$$

As before, in large samples the distribution of Y_i can be approximated by the normal with expected value and variance,

$$E(Y_i|H_1) = n_i\beta_i(\Psi_{i1} + 2\Psi_{i2}) + n_i(1 - \beta_i)(\pi_{i1} + 2\pi_{i2})$$

$$\text{Var}(Y_i|H_1) = n_i\beta_i[\Psi_{i1}(1 - \Psi_{i1}) + 4\Psi_{i2}(1 - \Psi_{i2}) - 4\Psi_{i1}\Psi_{i2}] + (1 - \beta_i)[\text{Var}(Y_i)|H_0] , \quad (13)$$

with $\text{Var}(Y_i|H_0)$ found in equation (5). Note that, in the case of a single gene and no phenocopies, the expected values and variances of Y_i under the normal approximation are found from equation (11) or equation (13), with β_i.

Alternative distribution of Z.—Under the alternative hypothesis, the exact density of Z at or near the kth disease locus is also a convolution and can be found from equation (6), by use of the appropriate distributions defined in equations (10) and (12). Finally, in large samples, the distribution of Z at or near the kth disease gene can be approximated by the normal with $E(Z) = \Sigma_{i=1}^{c} E(Y_i|H_1)$ and $\text{Var}(Z) = \Sigma_{i=1}^{c} \text{Var}(Y_i|H_1)$ and with use of the expected values and variances found from equations (11) and (13).

Alternative distribution of Z'.—In defining the alternative distribution of the largest order statistic, Z', it is useful to distinguish between the marker or clone that is at or near the kth disease locus and those that are unlinked to the disease locus. Thus we use $Z_{\theta \leq d}$ to denote the sum of the Y_i at recombination distance d from the disease locus, and we use $Z_{\theta=.5}$ to denote the sum of the Y_i for those markers or clones unlinked to disease locus k. The specified recombination frequency, d, is determined by marker spacing and a mapping function; or its value is 0 in the case of a GMS technology. With this notation, the alternative distribution of Z' in the set of markers or clones containing the disease gene is

$$F(Z':b) = P(Z' < b) = F(Z_{\theta=.5}:b)^{g-1}F(Z_{\theta \leq d}:b) , \quad (14)$$

with power $P(Z' \geq b) = 1 - F(Z':b)$. As in the null

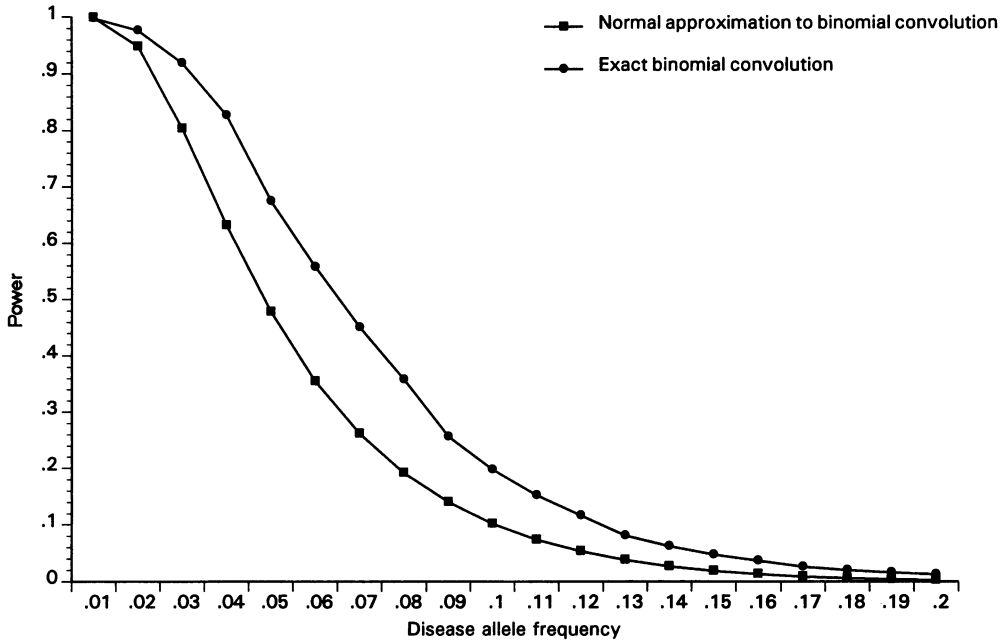

Figure 1 Power of the exact convolution and the normal approximation to detect a single autosomal codominant disease locus with a 10% population phenocopy rate in a sample of 75 sibling pairs using GMS (3,300 clones) ($\alpha_{cw} = .05$).

case, each Y_i is expressed in standard form, by use of the marker-specific mean and variance, which are a function of the number of alleles and the allele frequencies. Since we cannot be sure which marker will lie near a disease locus, we use the marker with the smallest PIC value to define the standardized Y_i for determining $Z_{\theta \leq d}$, yielding a conservative estimate of power.

For the case of a single susceptibility locus, a more precise definition of power in the current context is $P(Z_1 \leq Z_{\theta \leq d}, Z_2 \leq Z_{\theta \leq d}, \ldots, Z_{\theta \leq d} \geq b)$, the probability of the joint event that the largest order statistic equals or exceeds its critical value b and that the marker or clone corresponding to the largest order statistic is at or close to the kth susceptibility locus. For the exact test, a lower bound to this power is $F(Z_{\theta=.5}:b)^{g-1}[1 - F(Z_{\theta \leq d}:b)]$. For the large-sample normal test, this power is found from

$$\left(\Phi\left\{\frac{-[E(Z_{\theta=.5}) - E(Z_{\theta \leq d})]}{\sqrt{Var(Z_{\theta=.5}) + Var(Z_{\theta \leq d})}}\right\}\right)^{g-1} [1 - F(Z_{\theta \leq d}:b)] .$$

The preceding power is virtually the same as $1 - F(Z':b)$ at rare disease-allele frequencies, because the probability that a null marker or clone (i.e., $Z_{\theta=.5}$) exceeds by chance both the critical value b and $Z_{\theta \leq d}$ is negligible. The number of subjects required for a prespecified power can be computed exactly by use of the bisection algorithm, to invert any of the above expressions of power (Knuth 1973).

Comparisons of the exact and normal approximations of the power to detect a disease locus, across a range of genetic parameters and sampling designs considered in this paper, revealed inaccuracy in the normal approximation in small samples. In figure 1, we illustrate the power to detect a single codominant disease locus with a 10% population phenocopy rate in a sample of 75 sibling pairs, using GMS. In this example, when the disease allele frequency is .07, as many as 48 additional pairs of siblings are needed to compensate for the inaccuracy of the normal approximation. These results accentuate the importance of exact statistics in this context. Comparisons across other classes of relatives and models of genetic transmission show similar results in some cases (data not shown).

If there is more than one susceptibility gene, the alternative distribution of Z' may differ from equation (14), because the number of independent disease genes or their linked DNA segments in any one set is unknown. At one extreme, as described above, only a single gene may be present in a set of independent markers that are themselves unlinked to any other disease loci; at the other extreme, all the susceptibility loci may be, by chance, in the same set. For l genes in the same set of independent markers or clones,

$$F(Z':b) = F(Z_{\theta=.5}:b)^{g-1} \prod_{k=1}^{l} F_k(Z_{\theta \leq d}:b) .$$

Power under the single-locus case is a lower bound to that defined under the multilocus case; hence, for the multilocus models presented here, we use the lower bound.

Results

In the following sections, we compare the power of GMS and marker technologies to detect a susceptibility gene, across a range of models of inheritance and sampling strategies. Power curves are determined by use of either 330 equally spaced markers arranged into 10 sets of 33 markers or 3,300 clones arranged into 100 sets of 33 clones. Where appropriate, the Kosambi mapping function is used to determine θ, although use of any mapping function is possible. Power is computed by use of the normal approximation, unless otherwise stated.

For a clone-based GMS approach, we illustrate power to detect a susceptibility gene by using the procedure originally described by Nelson et al. (1993), and, in the case of siblings, we also evaluate MGMS (N. J. Schork and S. Ghosh, unpublished manuscript). For markers, we illustrate power under two heterozygosity values, .9 and 1. Heterozygosity of .9 is possible for a marker with 10 equally frequent alleles, and such markers are rapidly becoming available. Heterozygosity of 1 (i.e., I-IBD) may be achieved in several ways: (1) parental genotyping (e.g., in the sibling case), to identify unambiguous IBD status in sibling pairs; (2) development of extremely polymorphic markers; and (3) PGMS, i.e., hybridization of GMS-selected DNA to a subset of clones rather than to a complete set spanning the entire genome.

Etiologic Heterogeneity and Mixtures of Relative Classes

One of the most complex situations, probably representative of many human traits, is the presence of locus heterogeneity as well as nongenetic causes (phenocopies), in the pathophysiology of a disease trait. A general approach to reduce the inherent loss of power in the presence of etiologic heterogeneity is to obtain sufficiently large samples to enable identification of susceptibility genes that may be present only in a subgroup of affected pairs. In light of both likely etiologic heterogeneity and the need to identify large samples, it may be useful to combine different classes of relatives, to maximize sample-size needs. The approach presented here allows one both to calculate the proportion of pairs in each relative class that share the disease genotype at locus k (i.e., β_i), under a specified genetic model, and to compute power to find a disease gene in a mixed sample of independent pairs of affected relatives. In figure 2, we illustrate power to detect a susceptibility gene in a mixture of relative classes, composed of 100 sibling pairs, 75 GPGC pairs, and 75 first-cousin pairs, under a two-locus heterogeneity model with phenocopies, using GMS and marker technologies. The two disease alleles are assumed to be equally frequent, with an autosomal dominant mechanism of gene action. Values of x at locus 1 and locus 2 are equal and are computed at each disease-allele frequency to yield a population phenocopy rate of 10% at each locus.

The current approach allows comparison of the power of different search technologies, by use of a test with a common experiment-wise type I error probability. The three IBD approaches (GMS, GMS with MGMS for siblings, and I-IBD) have very similar power to detect a susceptibility gene under this mechanism of gene action and model of inheritance. The power of the GMS technology is slightly greater than the power of I-IBD at relatively rare allele frequencies; however, at an allele frequency of .08, the power of I-IBD surpasses that of GMS, and, at an allele frequency of .14, it surpasses the power of GMS with MGMS for siblings. A general comparison of these power curves, across molecular techniques, illustrates the dramatic gain in power by use of IBD information versus use of IBS information, even in the presence of highly polymorphic markers. When all three IBD states—rather than only two—are used with siblings, the gain in power is substantial, even in a sample composed of a mixture of relative pairs in which siblings constitute less than half the sample. More specific issues reflected in this figure are addressed in subsequent sections.

Etiologic heterogeneity.—In the presence of locus heterogeneity and/or phenocopies, power to detect a susceptibility gene is a function of both Ψ_i, i.e., the IBS probabilities, and β_i, i.e., the proportion of pairs in the sample of relatives of class i who both have the disease genotype at locus k. In table 1, we illustrate the effect of phenocopies and locus heterogeneity on β_i and power, under an autosomal dominant mechanism of inheritance and 150 GPGC pairs, using GMS. Recall that, in the case of GMS, $\mathbf{T} = \mathbf{D} = \mathbf{I}$; therefore, $\Psi_i = \lambda_i$. We vary the frequency of the disease allele at locus 1 while holding constant, at .05, the disease-allele frequency at locus 2. A population phenocopy rate of 10% is used at each locus, in the case of phenocopies. As shown in table 1, Ψ_{i1} remains constant in the presence of heterogeneity and/or phenocopies. In contrast, β_i is not constant, because it is affected by the presence of a second locus and/or phenocopies at each locus. There is relatively little loss in power to detect a susceptibility gene in the presence of a 10% phenocopy rate for a single gene when the disease-allele frequency is rare (i.e., <.1), but the loss can be substantial at a moderately common allele frequency (i.e., .2). This differential effect is a function of both Ψ_{i1} and β_i, which decrease as the kth disease allele becomes more common.

The presence of a second locus also has an effect on β_i, since the proportion of affected pairs that is due to locus k will increase in the presence of a less common, second disease allele. Obviously, when both loci have the same mechanism of inheritance and when the population allele frequencies are equal, $\beta_{i1} = \beta_{i2} = .5$. The simultaneous presence of locus heterogeneity and phenocopies further reduces β_i and power. As an example, in table 1, power is ~50%–80% less than that found under the single-locus case without phenocopies.

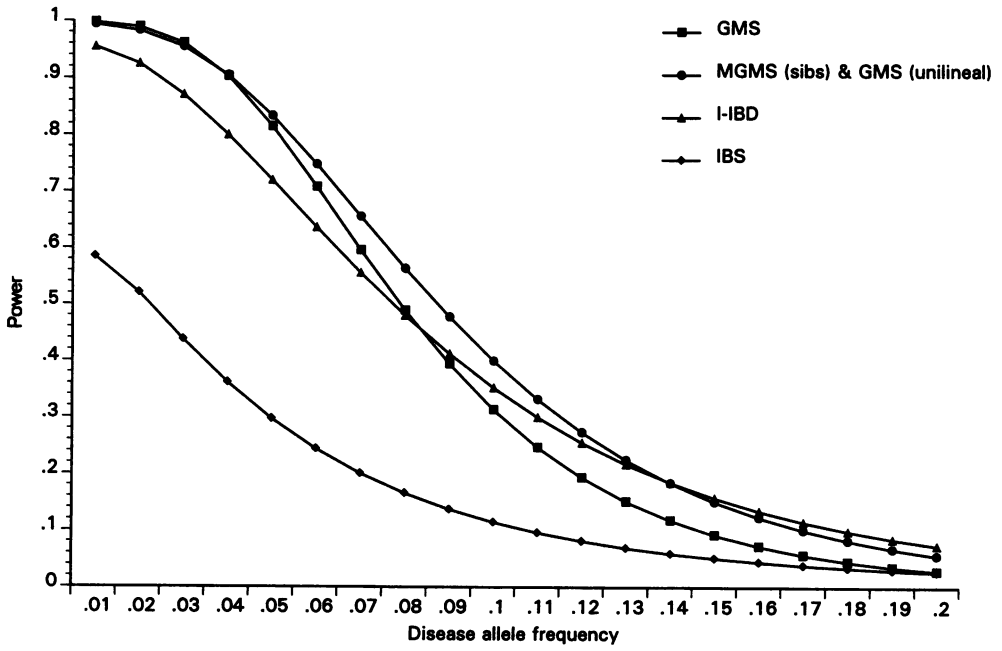

Figure 2 Power of GMS, GMS with MGMS modification for siblings, I-IBD, and IBS to detect a disease locus in a sample composed of 100 sibling pairs, 75 GPGC pairs, and 75 first-cousin pairs, under two-locus heterogeneity with two equally frequent autosomal dominant genes, each with a 10% population phenocopy rate (α_{ew} = .05). GMS technologies use 3,300 clones, and I-IBD and IBS use 330 equally spaced markers or clones (10 equally frequent alleles/marker for IBS) and the Kosambi mapping function (θ = .0498).

The impact of locus heterogeneity in the absence of phenocopies is illustrated in figure 3. In this figure, power to detect an autosomal dominant disease locus is given for 200 sibling pairs and 200 first-cousin pairs, under single-locus inheritance and two-locus heterogeneity with two equally frequent autosomal dominant genes, by use of I-IBD.

Several points can be inferred on the basis of this figure. First, power to detect a susceptibility gene in the presence of locus heterogeneity is much less than that found under single-gene inheritance, across both classes of relatives, particularly at common allele frequencies. However, there is also a differential effect of locus heterogeneity on power, across the classes of relatives, which becomes less dramatic as allele frequencies become more common. Specifically, at rare allele frequencies, power to detect a disease gene in a sample of first-cousin pairs is substantially less affected by locus heterogeneity than is such power in a sample composed of siblings.

Table 1

Effect of Locus Heterogeneity and Phenocopies on β_i and Power

		One-Locus Heterogeneity				Two-Locus Heterogeneity[a]			
Locus k		No Phenocopies		10% Population Phenocopy Rate		No Phenocopies		10% Population Phenocopy Rate/Locus	
Allele Frequence	Ψ_{i1}	β_i	Power	β_i	Power	β_i	Power	β_i	Power
.05	.85	1.00	1.00	.94	1.00	.50	.60	.48	.48
.10	.75	1.00	.99	.92	.96	.72	.63	.66	.48
.15	.69	1.00	.70	.89	.49	.82	.33	.74	.20
.20	.64	1.00	.22	.88	.10	.88	.09	.78	.04

NOTE.—Data are for a sample of 150 GPGC pairs, with use of GMS technology under autosomal dominant inheritance.

[a] The frequency of the disease allele at locus 2 is fixed at .05. Note that Ψ_{i2} = 0 in this example.

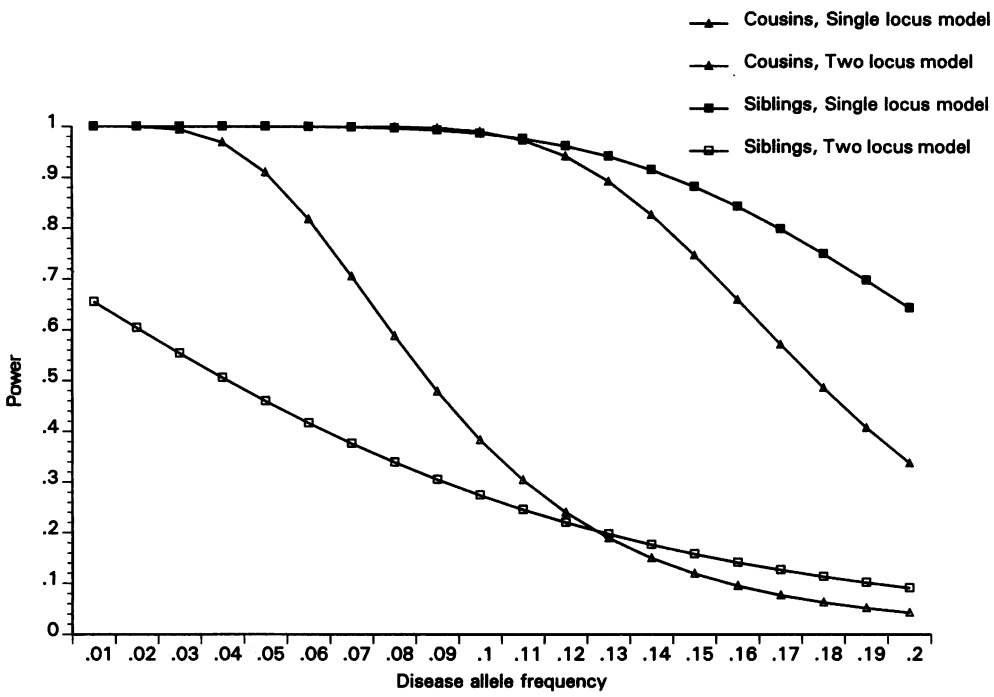

Figure 3 Power to detect an autosomal dominant disease locus with 200 sibling pairs and 200 first-cousin pairs under single-locus inheritance and two-locus heterogeneity with two equally frequent autosomal dominant genes ($\alpha_{ew} = .05$). Each curve contains I-IBD information based on 330 equally spaced markers or clones and the Kosambi mapping function ($\theta = .0498$).

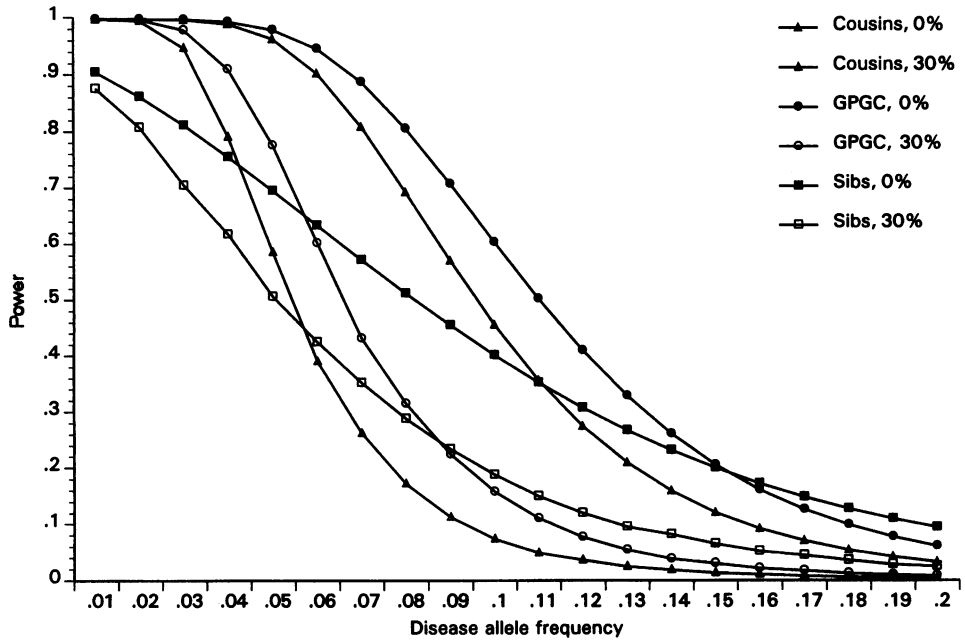

Figure 4 Power to detect an autosomal dominant disease locus in the presence of 0% and 30% population phenocopy rates in 75 sibling pairs, 75 GPGC pairs, and 75 first-cousin pairs ($\alpha_{ew} = .05$). Curves contain I-IBD information based on 330 equally spaced markers or clones and the Kosambi mapping function ($\theta = .0498$). An exact convolution is used for each curve.

The effect of phenocopies on power also can be isolated for any of the technologies described and for any mechanism and model of inheritance. As an illustration, in figure 4, we show the effect of a 30% population phenocopy rate under autosomal dominant single-locus inheritance in 75 sibling pairs, 75 GPGC pairs, and 75 first-cousin pairs, using I-IBD. An exact test is used for these comparisons, since sample sizes are small.

The effect of phenocopies on β_i and power reveals a different pattern across the classes of relatives than was seen for locus heterogeneity. For example, at a disease-allele frequency of .05, increasing the population phenocopy rate from 0 to 30% reduces power by 21% for GPGC, 27% for siblings, and 39% for first cousins. This differential effect across classes of relatives decreases as disease-allele frequencies become more common. Thus, for rare allele frequencies and autosomal dominant (or codominant) inheritance, power to detect a susceptibility locus is more sensitive to the presence of phenocopies in third-degree relatives than to its presence among first- or second-degree relatives. The implications for researchers designing a genome search is that, if locus heterogeneity and an autosomal dominant (or codominant) mechanism of gene action are suspected, one would achieve greater power by use of third-degree relatives than by use of first- or second-degree relatives, whereas, if phenocopies are the major concern, the reverse pattern would be true.

Mixtures of relative classes.—In addition to increasing the sample size, mixing the relative classes may be a useful sampling design, in light of the often unknown genetic etiology of a disease trait. Ideally, one would have some knowledge of likely mechanisms of gene action and mode of inheritance; however, in the presence of multiple loci or other complicating factors (e.g., phenocopies), it is quite plausible that an investigator's knowledge of the actual mechanism of inheritance might be incorrect at the outset of an investigation. Thus, another reason for combining classes of relatives might be the inherent lack of knowledge of gene action underlying a complex trait. Since various classes of relatives have differential power under different mechanisms of inheritance, a mixed sample of relatives may be more robust to misspecification of mechanisms of inheritance. In figure *5a* we show power to detect an autosomal recessive disease gene, and in figure *5b* we show an autosomal dominant disease gene under three sampling designs: 100 GPGC pairs, 100 sibling pairs, and a mixture of 50 GPGC and 50 sibling pairs, using I-IBD.

If a disease trait had a 20% population prevalence, then, under single-gene autosomal dominant inheritance, the corresponding disease-allele frequency would be .1, whereas under single-gene autosomal recessive inheritance this frequency would be .45. From figures *5a* and *5b,* it is clear that GPGC pairs are more powerful than siblings, under a dominant mechanism, but that the reverse is true under a recessive mechanism of gene action. By selecting the power curve associated with a combination of relative types, one will achieve sufficient power to detect a disease gene, under either of those two mechanisms of inheritance. Thus, by using a mixed sampling design, one could maximize the minimum sample size needed to achieve a specified power that is independent of the mechanism of inheritance.

Comparison of Technologies: GMS and Markers

We compare the power of the statistical test for GMS technology with that of markers, with experiment-wise $\alpha \leq .05$. For all comparisons, we compute power under single-gene autosomal dominant inheritance and in the absence of phenocopies. We first compare these technologies under the marker and clone strategies previously described for 330 markers and 3,300 clones, respectively, and then we evaluate alternative search strategies. In figure 6, we first compare the power of GMS, MGMS, I-IBD, and IBS to detect a single gene in 100 sibling pairs.

MGMS has greater power than either GMS or marker strategies, for the sibling sampling design, across the range of allele frequencies shown. Power to detect a susceptibility gene by use of a marker strategy, whether I-IBD or IBS, is greater than for that for GMS, at common allele frequencies. Together, these results demonstrate the difference between technologies that measure all three allele-sharing states (MGMS, markers) and one that measures only two states (GMS), and they support the development of MGMS to allow full discrimination of IBD status among siblings. A comparison of I-IBD and IBS curves illustrates that even a minor reduction in heterozygosity (from 1 to .9) leads to a substantial loss of power. Thus, the development of completely informative markers, whether through MGMS or other methods, will be critical for improving the power to find disease genes. Finally, comparison of the power of MGMS to that of I-IBD reveals that only a small gain in power is achieved by complete coverage of the genome ($\theta = 0$), a result that we subsequently will explore in more detail.

A comparison of marker and GMS technology for unilineal relatives (where MGMS is not applicable) is shown in figure 7. Power to detect a disease gene is shown for GMS, I-IBD, and IBS, in 100 GPGC pairs. Since the actual power to detect a susceptibility gene in a marker search depends on the distance of a marker to a disease locus, which could range from 0 cM to half the distance between adjacent markers, we present upper and lower bounds on power for the IBS case.

At rare allele frequencies GMS has greater power than the other approaches, and at more common frequencies it has power comparable to that of I-IBD. Thus, GMS

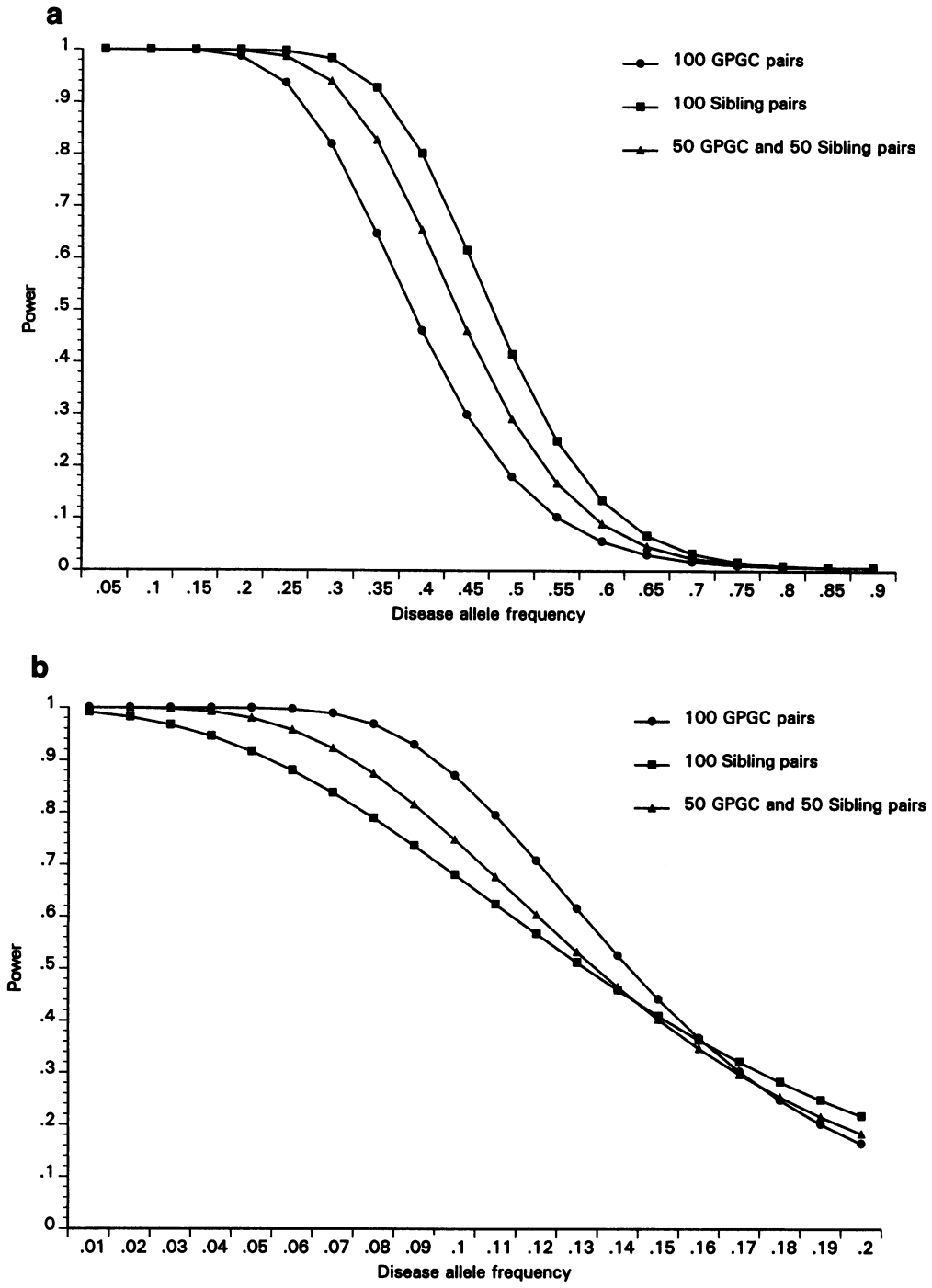

Figure 5 Power to detect an autosomal recessive disease locus (*a*) and an autosomal dominant disease locus (*b*) with 100 GPGC pairs, 100 sibling pairs, and a combination of 50 GPGC pairs and 50 sibling pairs (α_{ew} = .05). Each curve uses I-IBD information based on 330 equally spaced IBD markers or clones and the Kosambi mapping function (θ = .0498).

may be particularly useful for investigation of rare traits. The slight reduction in power of GMS compared with I-IBD, at common allele frequencies, reflects both the substantial increase in the number of test comparisons for GMS, compared with an interval search, and the more stringent per-comparison α required to achieve an equivalent experiment-wise α of .05. As is also shown in the figure, the power of an IBS marker approach does not exceed that of GMS, even under the best scenario of $\theta = 0$. Thus, consistent with results noted above, the development of completely informative markers through GMS or other methods is needed to increase the efficiency of a genome search.

We have seen that IBD approaches are by far more

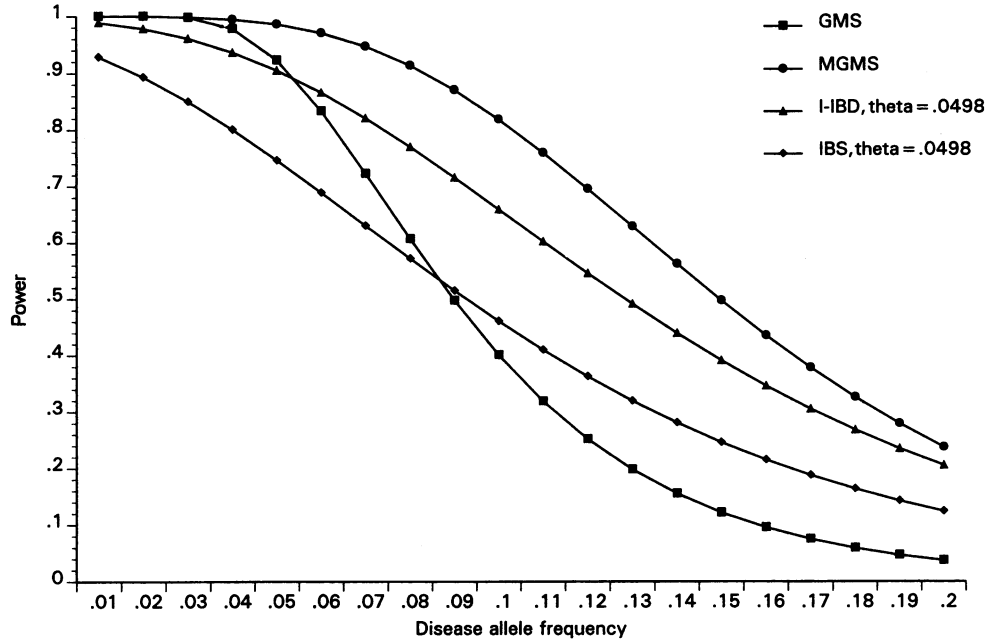

Figure 6 Power of GMS, MGMS, I-IBD, and IBS to detect an autosomal dominant disease locus with 100 sibling pairs (α_{ew} = .05). GMS and MGMS curves use 3,300 clones, and I-IBD and IBS curves use 330 equally spaced markers or clones (10 equally frequent alleles/marker for IBS) and the Kosambi mapping function (θ = .0498).

powerful than IBS, even at relatively high PIC values. However, obtaining IBD information from currently available marker technology is difficult. This problem has been approached in several ways, including maximum-likelihood estimation of IBD probabilities from IBS information (Risch 1990*c*), multipoint methods (Kruglyak and Lander 1995; Olson 1995), and use of certain relative pairs, particularly sibling pairs, in which IBD status can, in many cases, be unambiguously assigned through parental genotyping. Since parental ge-

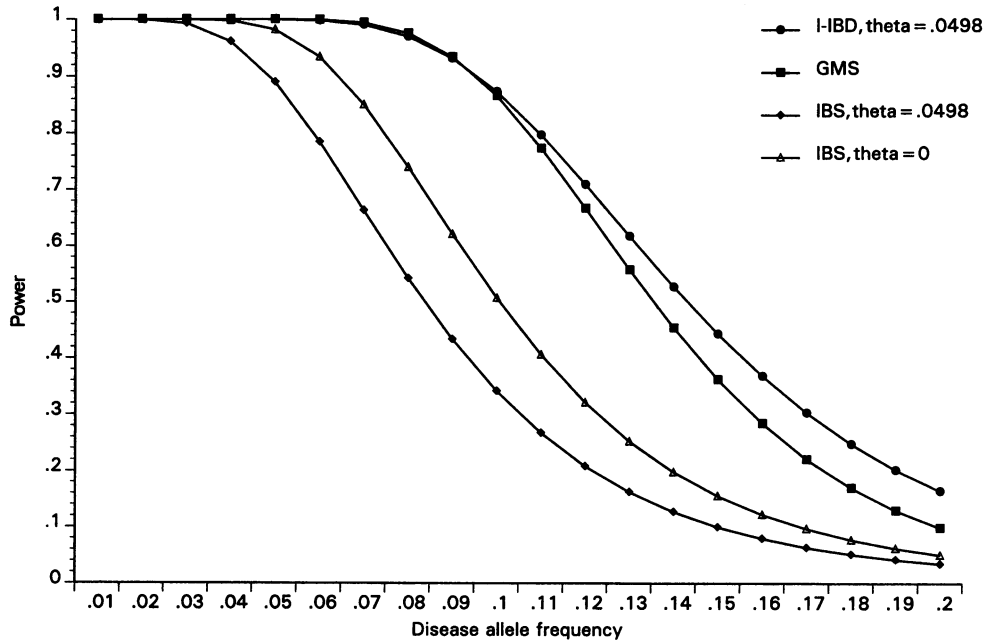

Figure 7 Power of GMS, I-IBD (θ = .0498), IBS (θ = .0498), and IBS (θ = 0) to detect an autosomal dominant disease locus with 100 GPGC pairs (α_{ew} = .05). GMS curve uses 3,300 clones, and I-IBD and IBS curves use 330 equally spaced markers or clones (10 equally frequent alleles/marker for IBS) and the Kosambi mapping function.

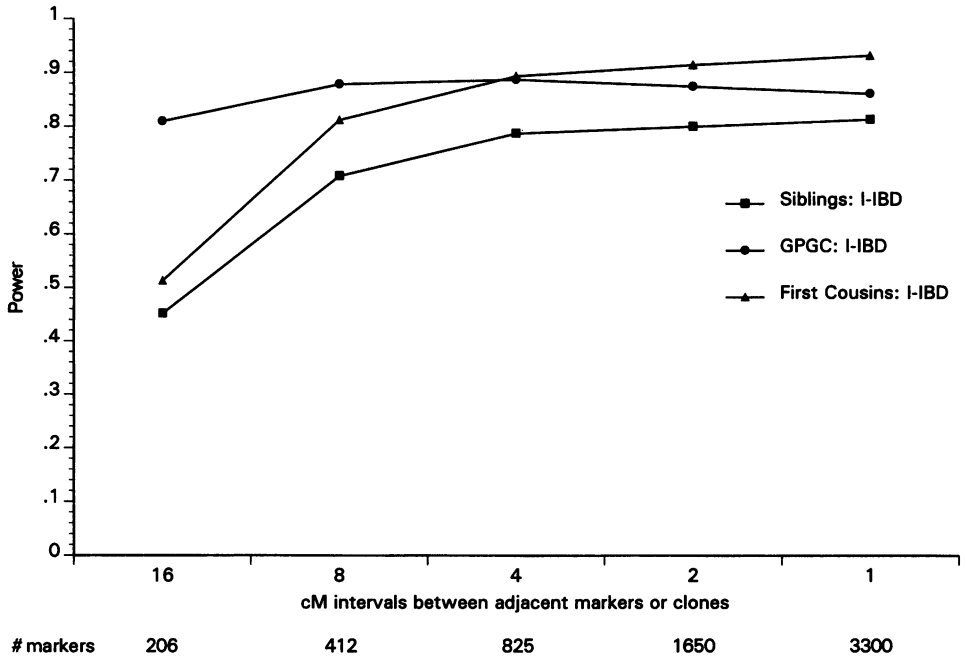

Figure 8 Power to detect a single autosomal dominant disease locus as a function of marker density (m = 206, 412, 825, 1,650, and 3,300) for I-IBD and IBS (10 equally frequent alleles/marker) (α_{ew} = .05). Power curves are shown for 100 sibling pairs, 100 GPGC pairs, and 100 first-cousin pairs at a disease-allele frequency of .1. The Kosambi mapping function is used.

notyping, required for the latter approach, involves an additional genotyping effort, compared with an IBS approach (which requires genotyping of sibling pairs only), we examined the sample sizes needed to achieve equivalent power under these two methods. To obtain 90% power under autosomal dominant single-gene inheritance with a disease-allele frequency of .1, an IBS method with marker heterozygosity of .9 would require genotyping 190 sibling pairs, or 380 persons. To obtain equivalent power by use of an IBD method, 177 sibling-pair families, or 708 persons (2 parents + 2 sibs per family) would require genotyping. This would allow for an expected exclusion of ~20% of the families, because of an inability to assign IBD status. Although, to obtain equivalent power, the IBS approach requires less genotyping than does the IBD approach, other factors make it less desirable; for example, findings are not robust to misspecification of allele frequencies (Babron et al. 1993). The substantial cost, in terms of subject sampling, to obtain IBD status in this way illustrates the importance of new technologies, such as GMS, in which IBD information can be directly obtained.

We have also seen that a GMS technology (including MGMS), which leads to complete genome coverage, has power similar to that of I-IBD methods with 10-cM spacing between adjacent markers or clones. We now address the question, What marker or clone density is needed to obtain optimal power in a genome search? In figure 8 we illustrate the power of an I-IBD search, as a function of density of the markers or clones (i.e., θ) for three classes of relatives, each of size 100, at a disease-allele frequency of .1. Power significantly increases until the marker or clone density approaches an 8-cM spacing (θ = .04), for all three relative classes. At that point, power nominally increases as marker or clone density increases. For GPGC, however, power begins to decline as marker or clone density increases beyond 4-cM spacing, a finding also seen at different allele frequencies and under autosomal recessive inheritance. The change in power as the number of markers increases is a consequence of both the increased number of test comparisons (hence, a more stringent per-comparison α) and the decreased θ. These results suggest that adequate power can be achieved by a PGMS procedure in which approximately one-eighth of the human genome is used for hybridization. These data further suggest that, as marker genotype scoring increases in efficiency, adequate power to detect a locus can be achieved by spacing markers at ~8-cM intervals. A genome search using more closely spaced markers achieves no substantive increase in power.

Discussion

The availability of a dense set of polymorphic markers and novel techniques such as GMS are making it feasible to scan the human genome to detect and localize susceptibility genes even when there is little a priori information regarding the genetic etiology of a trait. However, if a statistical test that does not explicitly control for

the possibility of no disease gene is applied in such an exploratory genome search, it will lead to an elevated probability of false detection of linkage.

We have presented a general statistical approach that is appropriate for samples of independent pairs of affected relatives, that can be used with either marker or GMS technologies, and that has an exact test. The approach can be used with any mixture of relative classes, any number of markers or clones, and any mapping function. In the presence of etiologic heterogeneity, both the number of pairs in which both members have the disease genotype, $n_i\beta_i$, and the number of pairs in which at least one member of a pair is a phenocopy or is affected because of a different locus, $n_i(1 - \beta_i)$, are calculated under a specified genetic model and are used in the computation of power. As with other affected-pair methods, we assume accurate specification of marker-allele frequencies for IBS data, and we assume linkage equilibrium between a susceptibility gene and an adjacent marker or clone. In addition, for clone-based GMS, we assume that hybridization signals of 1-Mb-size clones are read without error. This will, in reality, depend on the number of GMS fragments that hybridize to the clone, a process that will be a function of both the size of GMS-selected fragments and other factors. Further work is needed to incorporate a measurement model into the statistical framework.

The approach presented here was inspired by the elegant work of Feingold (1993) and Feingold et al. (1993), who, anticipating the emergence of GMS, presented a statistical test for localizing disease genes by using this technology that controls for the type I error. By using an order statistic and Gaussian approximation for the sum of dependent Bernoulli processes, they defined power to detect a disease locus on a chromosome, under different models of inheritance and different sampling designs. Using independent sets of markers or clones, we have extended the scope of issues considered by their approach, by incorporating both GMS and marker technologies, etiologic complexity, and exact-test statistics.

As currently formulated in this paper, power to detect a single locus is more conservative than is power defined by Feingold et al. (1993). We propose taking enough subjects to have a prespecified power to detect the disease gene itself. In the Feingold et al. (1993) definition of power, one would take enough subjects to detect a disease gene or any linked DNA segment on the same chromosome. Nevertheless, the power of the normal approximation presented here is greater than the normal approximation of Feingold et al., in some cases. For example, with a single codominant gene, 100 sibling pairs, and GMS, power under the Feingold normal approximation is .902 and .970 for disease-allele frequencies of .04 and .01, respectively, whereas the respective powers under the normal approximation of this paper are .981 and .999. At more common allele frequencies, the power of the Feingold et al. normal approximation can exceed the power of the normal approximation presented here. For example, at allele frequencies of .076 and .2, power is .754 and .240 under the Feingold et al. (1993) normal approximation, whereas it is .696 and .048 under the normal approximation presented in this paper.

Although our definition of power is conservative relative to that of Feingold et al., our approach has the advantage of having an exact test. As shown previously, the difference, in power, between a normal approximation and an exact test may in some cases be substantial, especially in small samples. Even when the overall sample size is large, the presence of etiologic heterogeneity can lead to a small subgroup of pairs in which at least one member is a phenocopy or is affected because of a different disease locus.

Finally, the model can be used either to derive practical information for designing studies of complex traits or to examine differences in performance of existing and emerging technologies. First, we have found that the effect of etiologic heterogeneity on power depends on the source of the heterogeneity and on the class of relatives under investigation. Specifically, under an autosomal dominant mechanism of gene action, third-degree relatives provide greater power than first- or second-degree relatives, in the presence of locus heterogeneity, whereas the reverse is true when phenocopies are the relevant issue. Second, when the mode of inheritance is unknown, a mixture of relative classes may be more robust to misspecification of mechanisms of inheritance and will maximize the minimum sample size needed to achieve a specified power. By comparing the performance of different technologies, our results have revealed a finding particularly relevant to the development and application of GMS. Specifically, a partial GMS procedure, PGMS, in which only one-eighth of the genome is used, has approximately the same power as does hybridization to the complete genome. Thus, use of a procedure that obtains IBD information from partial—rather than complete—genome coverage will result in no appreciable loss in power. Furthermore, even if GMS can successfully isolate IBD information from the entire human genome, the subsequent evaluation process could be simplified by hybridization to 88% fewer clones.

Acknowledgments

This work was supported by grants from NIMH (MH46981 and MH44742, both to S.S.) and from NIDA (DA01070 to J.A.W.). We wish to thank Stanley Nelson for his contributions in descriptions of GMS and for critical review of earlier manuscripts, and we wish to thank Eleanor Feingold for providing us with specific power calculations and with clarification of issues regarding use of her model. We also wish to thank an anonymous reviewer for suggesting an efficient algorithm for computing the exact convolution with discrete random vari-

ables. Also, we thank Kathy Kim for her help in manuscript preparation.

Appendix A

Computation of the density function of the sum of independent discrete random variables having unequal probabilities of allele sharing is required for the exact tests of this paper. The density function can be computed by use of a discrete convolution algorithm such as those found in the work of Feller (1957) or Press et al. (1992). However, these convolution algorithms are not computationally feasible for the genetic applications considered in this paper. A convolution algorithm that is computationally feasible in this context was suggested by an anonymous reviewer. Given the sum of independent discrete random variables with limited range, say (0,1,2), denoted as $Y_n = X_1 + X_2 + \ldots + X_n$, the distribution of Y_{n+1} equals

$$\begin{aligned} P(Y_{n+1} = i) = {} & P(Y_n = i)P(X_{n+1} = 0) \\ & + P(Y_n = i - 1)P(X_{n+1} = 1) \\ & + P(Y_n = i - 2)P(X_{n+1} = 2) , \end{aligned}$$

where $P(Y_n = i)$ is known and is 0 if i falls outside the range of that random variable.

For MGMS or siblings with markers, the density of equation (4) is computed by taking Y_n to be the first three-level discrete random variable with $P(Y_n = i)$ equal to the elements of π_i. The remaining random variables for that relative class are added one at a time. For summing across relative classes with GMS or unilineal relatives with markers, the convolution of binomials can be computed more efficiently by taking Y_n to be the binomial random variable for the first class and adding one at a time the Bernoulli random variables from the other relative classes. In the null case of MGMS and siblings, the distribution of the sum of the three-level independent random variable given in equation (4) can be conveniently computed as Bin($2n$,.5) as shown by Green and Woodrow (1977).

Appendix B

In this appendix we present expressions needed to define β_i, the proportion of affected pairs in a sample, both of whom have disease gene k, that is appropriate for any mechanism of gene action, phenocopies, and locus heterogeneity and for any class of relatives. From these expressions, we can define β_i and $(1 - \beta_i)$, the latter being the proportion of pairs in which at least one member is a phenocopy or at least one member is affected because of a different disease locus.

Under a single-gene mode of inheritance, $P(2\text{ AR})$ for relative class i, denoted γ_i, is defined on the basis of the allele frequencies at the disease locus, the penetrances, and the probabilities of IBD associated with the degree of genetic relatedness, κ_i. We assume that the disease gene has two alleles with frequencies p and q, where p represents the disease allele frequency.

The following four matrices are used to define γ_i:

$$\mathbf{t}' = [1 \quad a \quad x] \qquad \mathbf{g}' = [p^2 \quad 2pq \quad q^2]$$

$$\mathbf{G} = \begin{bmatrix} p^3 & p^2q & 0 \\ p^2q & pq & pq^2 \\ 0 & pq^2 & q^3 \end{bmatrix} \quad \mathbf{MZ} = \begin{bmatrix} p^2 & 0 & 0 \\ 0 & 2pq & 0 \\ 0 & 0 & q^2 \end{bmatrix} . \tag{B1}$$

Under the assumption of random mating, the genotype probabilities under Hardy-Weinberg are given in matrix **g**. Penetrances are defined in the **t** vector, where dominant, codominant, and recessive inheritance are modeled by $a = 1$, $1/2$, and x, respectively. The element x represents the relative penetrance of individuals who do not carry the disease allele compared with those who do, according to the method of Bishop and Williamson (1990). Thus, the absence of phenocopies is represented by $x = 0$, and the presence of phenocopies is represented by $x > 0$ (but understood to be ≤ 1). **G** is the joint genotype-probability matrix for parent-offspring, and **MZ** is the joint-genotype-probability matrix for MZ twins.

We next define three conditional probabilities with $x \geq 0$:

$$P(2\text{ AR}|\text{IBD} = 0) = \mathrm{A} = [\mathbf{1}'(\mathbf{g} \times \mathbf{t})]^2 ,$$

$$P(2\text{ AR}|\text{IBD} = 1) = \mathrm{B} = \{\mathbf{1}'[\mathbf{G} \times (\mathbf{tt}')]\mathbf{1}\} ,$$

$$P(2\text{ AR}|\text{IBD} = 2) = \mathrm{C} = \{\mathbf{1}'[\mathbf{MZ} \times (\mathbf{tt}')]\mathbf{1}\} ,$$

where $[\ ] \times [\ ]$ indicates the hadamard product of matrices. On the basis of these probabilities and the known coefficients of relationship, $P(2\text{ AR})$ is

$$\gamma_i = [\mathrm{A}_{x\geq 0} \quad \mathrm{B}_{x\geq 0} \quad \mathrm{C}_{x\geq 0}] \begin{bmatrix} \kappa_{i0} \\ \kappa_{i1} \\ \kappa_{i2} \end{bmatrix} .$$

By restricting $x = 0$, we can define $P(2\text{ AR at locus } k)$ as

$$\gamma_i^* = [\mathrm{A}_{x=0} \quad \mathrm{B}_{x=0} \quad \mathrm{C}_{x=0}] \begin{bmatrix} \kappa_{i0} \\ \kappa_{i1} \\ \kappa_{i2} \end{bmatrix} .$$

Table B1

D_i Matrix: Conditional IBD Matrix for Two Linked Loci r and k for Relative Class i

Relative Class i	IBD at Marker Locus r	IBD at Trait Locus k		
		0	1	2
Sibs	0	ω^2 [a]	$2\omega(1-\omega)$	$(1-\omega)^2$
	1	$\omega(1-\omega)$	$\omega^2+(1-\omega)^2$	$\omega(1-\omega)$
	2	$(1-\omega)^2$	$2\omega(1-\omega)$	ω^2
Grandparent-grandchild	0	$(1-\theta)$	θ	0
	1	θ	$(1-\theta)$	0
	2	0	0	0
Aunt-niece	0	$\omega(1-\theta)+\frac{1}{2}\theta$	$1-\frac{1}{2}\theta-\omega(1-\theta)$	0
	1	$1-\frac{1}{2}\theta-\omega(1-\theta)$	$\omega(1-\theta)+\frac{1}{2}\theta$	0
	2	0	0	0
Half sibs	0	ω	$(1-\omega)$	0
	1	$(1-\omega)$	ω	0
	2	0	0	0
First cousins	0	$\frac{1}{3}[2+\frac{1}{2}\theta^2+\omega(1-\theta)^2]$	$\frac{1}{3}[1-\frac{1}{2}\theta^2-\omega(1-\theta)^2]$	0
	1	$1-\frac{1}{2}\theta^2-\omega(1-\theta)^2$	$\omega(1-\theta)^2+\frac{1}{2}\theta^2$	0
	2	0	0	0

[a] $\omega=\theta^2+(1-\theta)^2$.

In γ_i the x in the penetrance vector **t** is free, and in γ_i^*, x is fixed at zero.

In general, β_i is a function of γ_i and γ_i^*. As an example, for single-gene inheritance with phenocopies, $\beta_i = \gamma_i^*/\gamma_i$, the proportion of affected pairs of relative class i in which both members carry the disease gene k; that is, neither member is a phenocopy. We can also define β_i in the presence of l disease loci under locus heterogeneity and epistasis, with or without phenocopies. In the presence of more than one disease locus, the matrices in equation (B1) can be defined for each of l unlinked disease loci, where $l = 1,k$. The $P(2\text{ AR})$, i.e., γ_i, is a function of the between-locus interaction, which can be modeled through penetrance functions, either as the union of individual penetrances (i.e., heterogeneity), according to Risch (1990*a*, p. 225; Genetic Heterogeneity Model), or as the product (i.e., epistasis), according to Hodge (1981) and Risch (1990*a*). For convenience, we introduce the subscript k in the following definitions of β_i, to differentiate among the l loci.

For the kth locus among a set of l heterogeneous loci,

$$\beta_i = \frac{\gamma_{ik}^*}{\sum_{k=1}^{l}\gamma_{ik} - \Sigma\Sigma\,\gamma_{ik}\gamma_{im} + \Sigma\Sigma\Sigma\,\gamma_{ik}\gamma_{im}\gamma_{in} - , \ldots, + (-1)^{l+1}\prod_{k=1}^{l}\gamma_{ik}}.$$

Notice that, when $l = 1$, β_i reduces to that shown previously for the single-gene case in the presence of phenocopies.

For the kth locus among a set of l epistatic loci,

$$\beta_i = \frac{\gamma_{ik}^*\gamma_{im}\cdots\gamma_{il}}{\gamma_{ik}\gamma_{im}\cdots\gamma_{il}} = \frac{\gamma_{ik}^*}{\gamma_{ik}}.$$

In this case β_i is equivalent to that shown for the single-locus case, since, under epistasis, all loci are necessary for determination of the trait.

Using the notation of the present paper, we define λ_i, the vector of conditional IBD probabilities for the ith relative class holding $x = 0$ at the disease locus k, as

$$\boldsymbol{\lambda}_i = \begin{bmatrix}\lambda_{i0}\\ \lambda_{i1}\\ \lambda_{i2}\end{bmatrix} = \begin{bmatrix}(A_{x=0}\kappa_{i0})/\gamma_i^*\\ (B_{x=0}\kappa_{i1})/\gamma_i^*\\ (C_{x=0}\kappa_{i2})/\gamma_i^*\end{bmatrix}. \qquad \text{(B2)}$$

Definition of Ψ_i, the IBS probabilities at a marker or clone r linked to disease locus k at recombination frequency θ, can be defined analogously with $x = 0$, through equation (9), by use of $\boldsymbol{\lambda}_i$, T, and D_i. The elements of the matrix T have been defined by Bishop and Williamson (1990, p. 255), $\boldsymbol{\lambda}_i$ is defined in equation (B2), and D_i is provided in table B1, in terms of the notation used in this paper.

References

Babron M-C, Martinez M, Bonaiti-Pellie C, Clerget-Darpoux F (1993) Linkage detection by the affected-pedigree-member method: what is really tested? Genet Epidemiol 10:389–394

Bishop DT, Williamson JA (1990) The power of identity-by-state methods for linkage analysis. Am J Hum Genet 46:254–265

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85–97

Brown DL, Gorin MB, Weeks DE (1994) Efficient strategies for genomic searching using the affected-pedigree-member method of linkage analysis. Am J Hum Genet 54:544–552

Elston RC (1992) Designs for the global search of the human genome by linkage analysis. Proceedings of the International Symposium on Human Genetics and Variation & XVIII Annual Conference of the Indian Society of Human Genetics, Hyderabad, December 11–13

Feingold E (1993) Modeling a new genetic mapping method. Tech rep 11, Department of Statistics, Stanford University, Palo Alto

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234–251

Feller W (1957) An introduction to probability theory and its applications, 2d ed. John Wiley & Sons, New York

Green JR, Woodrow JC (1977) Sibling method for detecting HLA-linked genes in disease. Tissue Antigens 9:31–35

Hochberg V, Tamhane AC (eds) (1987) Multiple comparison procedures. John Wiley & Sons, New York

Hodge DE (1981) Some epistatic two-locus models of disease. I. Relative risks and identity-by-descent distributions in affected sib pairs. Am J Hum Genet 33:381–395

Knuth DE (1973) Sorting and searching. Addison-Wesley, Reading, MA

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

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

Lange K (1986) The affected sib-pair method using identity by state relations. Am J Hum Genet 39:148–150

Nelson SF, McCusker JH, Sander MA, Kee Y, Modrich P, Brown PO (1993) Genomic mismatch scanning: a new approach to genetic linkage mapping. Nat Genet 4:11–13

Olson JM (1995) Multipoint linkage analysis using sib pairs: an interval mapping approach for dichotomous outcomes. Am J Hum Genet 56:788–798

Ott J (1991) Analysis of human genetic linkage. Johns Hopkins University, Baltimore

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in FORTRAN, 2d ed. Cambridge University Press, New York

Renwick JH (1969) Progress in mapping human autosomes. Br Med Bull 25:65–73

Risch N (1990*a*) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228

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

——— (1990*c*) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253

——— (1991) A note on multiple testing procedures in linkage analysis. Am J Hum Genet 48:1058–1064

Suarez BK (1978) The affected sib pair IBD distribution for HLA-linked disease susceptibility genes. Tissue Antigens 12:87–93

Thomson G, Motro U (1994) Affected sib pair identity by state analyses. Genet Epidemiol 11:353–364

Wright S (1922) Coefficients of inbreeding and relationship. Am Nat 56:330–338