*Am. J. Hum. Genet. 63:225–240, 1998*

# Mapping of Complex Traits by Single-Nucleotide Polymorphisms

Lue Ping Zhao, Corinne Aragaki, Li Hsu and Filemon Quiaoit

Quantitative Genetic Epidemiology Group, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle

## Summary

**Molecular geneticists are developing the third-generation human genome map with single-nucleotide polymorphisms (SNPs), which can be assayed via chip-based microarrays. One use of these SNP markers is the ability to locate loci that may be responsible for complex traits, via linkage/linkage-disequilibrium analysis. In this communication, we describe a semiparametric method for combined linkage/linkage-disequilibrium analysis using SNP markers. Asymptotic results are obtained for the estimated parameters, and the finite-sample properties are evaluated via a simulation study. We also applied this technique to a simulated genome-scan experiment for mapping a complex trait with two major genes. This experiment shows that separate linkage and linkage-disequilibrium analyses correctly detected the signals of both major genes; but the rates of false-positive signals seem high. When linkage and linkage-disequilibrium signals were combined, the analysis yielded much stronger and clearer signals for the presence of two major genes than did two separate analyses.**

## Introduction

Technological advances in molecular genetics have been a driving force in the rapid progress of the Human Genome Project and modern human genetics. Shortly after the construction of the first-generation human genome map on the basis of RFLPs, researchers developed the second-generation human genome map, using microsatellite markers (Donis-Keller et al. 1987; Murray et al. 1994). Following these successes, molecular geneticists are constructing another human genome map, with single-nucleotide polymorphisms (SNPs), using chip-based microarrays (Chee et al. 1996; Lipshutz 1997; Wang et al. 1998). The first commercial SNP chip is expected to contain 2,000 SNP loci, which should be superior, in information content, to current microsatellite-marker sets (Kruglyak 1997). Although the density of SNP loci varies across the genome, the average density is a~2–3 cM. Among other desirable properties, chip-based microarrays promise high throughput economically. This technology will have many uses, including an application to the mapping of complex traits.

Taking full advantage of the biallelic nature of SNP markers and of their high density, we introduce a semiparametric method for the mapping of complex traits via linkage and linkage-disequilibrium analysis. To improve the efficiency of linkage analysis, this method pools several adjacent SNP loci and estimates an averaged recombination fraction (RF), with use of marker loci. This averaged estimate may be interpreted as an RF with a small genome segment of those SNP markers. Meanwhile, the method estimates parameters quantifying linkage disequilibrium, which is likely, in view of the high density of SNP markers. Furthermore, combining both linkage and linkage-disequilibrium analysis, this method produces a combined test statistic that allows one to detect the presence of both linkage and linkage disequilibrium. This combined test generally reduces the false-positive errors produced by separate analyses, especially by linkage analysis.

Using this semiparametric approach, we analyze a set of nuclear-family data simulated in a genome-scan experiment. This experiment simulates two major genes that are 100 cM apart on a map of 44 SNP markers; one of these genes confers a high penetrance with low allele frequency, and the other gene confers a modest penetrance and modest allele frequency. Analyzing the simulated genome-scan data, we have found that linkage analysis has correctly identified two major genes and yet seems to experience high false-positive errors. On the other hand, linkage-disequilibrium analysis appears to capture signals of both major genes. The signals become much clearer when linkage analysis and linkage-disequilibrium analysis are combined. The limited experience with this simulated genome-scan experiment indicates the potential utility of SNP markers and of this semiparametric method in the mapping of complex traits.

Received December 2, 1997; accepted for publication May 1, 1998; electronically published June 19, 1998.

Address for correspondence and reprints: Dr. Lue Ping Zhao, Quantitative Genetic Epidemiology, MW806, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109. E-mail: lzhao@fhcrc.org


0002-9297/98/6301-0000$02.00

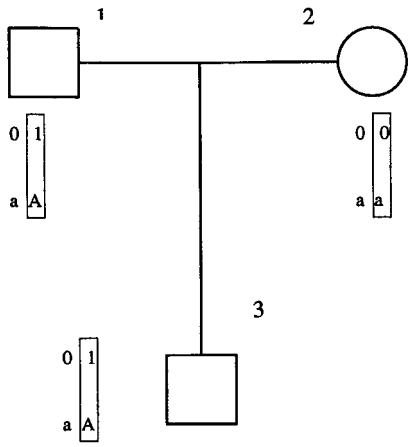


**Figure 1** Sample pedigree, illustrating the phenomenon of linkage between an SNP marker locus and the putative disease locus and that of linkage disequilibrium between SNP allele and the putative disease allele.

## Material and Methods

### *Biallelic SNPs*

The human genome is full of SNPs; it is estimated that there is, on average, one SNP locus for every 500–1,000 nucleotides (Wang et al. 1998). These SNPs can be used as "signposts" for the human genome. Each SNP locus is commonly expected to be biallelic and polymorphic. Notationally, let $m_k = m_{k1}m_{k2}$ denote a pair of alleles, forming an SNP genotype at the $k$th locus, where $m_{k1}(m_{k2}) = 1$ or $m_{k1}(m_{k2}) = 0$, corresponding, respectively, to the presence or absence of the designated allele (either one of A, C, G, or T), and where $k =$ 1,2,...,2,000 for 2,000 SNP loci on the Poly2000 GeneChip.

### *An Introduction to Linkage/Linkage-Disequilibrium Analysis*

For simplicity, consider one SNP marker and one putative disease locus. Suppose that the disease locus has deleterious allele "A" and normal allele "a." Furthermore, consider a nuclear family with two parents and a son (fig. 1), in which the mother is homozygous at both the marker locus and the putative disease locus, with respective genotypes (0|0) and (a|a). She passes the haplotype "0a" to her son. Suppose that the father is heterozygous at both the marker locus and the putative disease locus, with known haplotypes "0a" and "1A," respectively. In the absence of recombination for the paternal meiosis, either 0a or 1A will be passed to his son; otherwise, in the presence of recombination, either 0A or 1a will be passed to his son.

*Linkage Analysis and RF.*—The primary objective of a linkage analysis is to estimate the RF, which can be used to determine the genetic distance between the putative disease locus and the marker locus. The RF is simply a fraction of recombinants, of all informative meioses. For example, for data given in figure 1, one can directly observe an informative paternal meiosis that is a nonrecombinant. Let $y = 1$ or $y = 0$denote, respectively, presence or absence of recombination in the informative meiosis. Statistically, linkage analysis is equivalent to estimating the mean of the recombination—that is, $\theta =$ E($y$). If the estimated RF is significantly $<.5$, then one would declare a positive linkage.

*Linkage-Disequilibrium Analysis and the Odds Ratio (OR).*—The presence of linkage disequilibrium implies that the disease allele at the putative disease locus is associated with an allele at the marker locus. Linkage disequilibrium is equivalent to the association between the putative disease allele $g$ and the marker allele $m$; that is, they are no longer independent. The allele distribution for the parental haplotypes in figure 1 is represented in table 1. One way of capturing the empirical association between two binary alleles is by use of the OR, via a logistic regression of the marker allele, given the putative disease allele; this may be written as

$$P(m = 1|g) = 1/[1 + \exp(\alpha + \beta g)] , \qquad (1)$$

where $\exp(\beta)$ is the OR and, hence, $\beta$ is referred to as the log OR. If $\beta$ is significantly different from 0, formula (1) indicates that $m$ and $g$ are dependent—that is, in linkage disequilibrium. The intercept $\alpha$ also has an intuitive interpretation under linkage equilibrium, via the marker-allele frequency, $P(m = 1|g) = P(m = 1) = 1/[1 + \exp(\alpha)]$. Note that we choose to model $P(m = 1|g)$, instead of $P(g = 1|m)$, because marker data in mapping studies are generally random and because genotypes, indirectly, via phenotypes, are subject to ascertainment.

### *An Introduction to Multipoint Linkage Analysis*

A typical mapping study with multiple markers would generally perform a multipoint linkage analysis. The multipoint linkage analysis is known to be efficient, because it incorporates the map information—that is, the order and distances between markers. Specifically, suppose that $M$ marker loci are ordered linearly with known

**Table 1**

**Allele Distribution, for Putative Genotype and Marker Genotype**

| Allele at Putative Disease Locus | Distribution of Marker Allele | |
|---|---|---|
| | 0 | 1 |
| a | $n_{0a} = 3$ | $n_{1a} = 0$ |
| A | $n_{1A} = 0$ | $n_{1A} = 1$ |

order and known genetic distances; for example, $\delta_1 = 0 - \delta_2 - \delta_3 - \cdots - \delta_M$ is the map of $M$ loci, treating the far-left locus as the reference, where $\delta_l$ is the distance of the $l$th marker from the reference point $\delta$. Let $\delta_x$ denote the map position of the putative disease locus. Then, the RFs of the putative disease locus with all $M$ loci are functions of $\delta_x$; that is, $\theta_{xl} = h(|\delta_x - \delta_l|)$, where $h(\cdot)$ is a chosen mapping function—for example, that of Haldane (1919). While appreciating the efficiency gain by such a multipoint linkage analysis, one should also realize that this gain also benefits from the assumption that there are known genetic distances and additivity of map distances. Moreover, these predicted RFs from the map may be incomparable with estimated RFs, because the latter are generally biased toward null, owing to the misspecification of the segregation models.

An alternative to be considered here is the use of averaged RFs of several loci adjacent to the locus of interest. Specifically, for each locus, one can identify its neighboring loci and then can estimate an averaged RF between the putative disease locus and several adjacent marker loci. In essence, one is smoothing the estimated RF for the $l$th locus, by borrowing its neighboring information, just as the nonparametric estimation of a function by a smoothing technique (e.g., see Silverman 1990) . Let $\bar{\theta}_l$ denote the averaged RF between the $l$th locus and those loci within its neighborhood. Notationally, $\bar{\theta}_l = \frac{1}{c}\Sigma_{k\in N_l}\theta_k$, where the summation is over all $c$ loci in the neighborhood ($N_l$) of the $l$th locus. The neighborhood can be defined as those loci within a specified map distance, if mapping distances of all marker loci are known. Even if these distances are unknown or are not trustworthy, the neighborhood can be defined by the adjacency of marker loci, provided that the order of all marker loci is unambiguously specified. In the case of mapping studies with SNP markers, the neighborhoods can be defined by use of either map distances or the order of markers, because SNP markers are chosen from well-established genetic-marker regions, as well as because the map density is expected to be high. As expected, the averaged estimate has an improved efficiency, which may be interpreted as pooling several SNP loci to achieve the efficiency attained by microsatellite markers. This pooling strategy may become ineffective, however, if the chosen neighborhood of the $l$th locus is too wide to be specific. In simulation studies to be described, two SNP loci are pooled.

### *Practical Difficulties*

For our purpose, linkage/linkage-disequilibrium analysis may be thought of as estimating RF for linkage and as estimating log OR for linkage disequilibrium. However, in human studies, the analysis will encounter several practical difficulties. First of all, the putative disease alleles are generally unobserved and need to be inferred on the basis of the observed phenotypes and their familial aggregation. Second, even if putative disease allele is inferred to be heterozygous, the parental origins of putative disease alleles among all family members are unknown and hence must be numerated in the calculation. Third, the parental origins of marker alleles among family members may be partially determined. Consequently, haplotyping information may not be available. Fourth, to overcome the loss of heterozygosity with SNP markers, one must consider multipoint linkage analysis to improve linkage-analysis efficiency, but the computation is a challenge.

### *Notation*

Consider a family study that ascertains $I$ ($i = 1, \ldots, I$) families, each of which includes $n_i$ ($j = 1, \ldots, n_i$) family members. For the $j$ th member in the $i$th family, the phenotype, SNP markers, and other covariates are collected. Let $d_{ij}$ denote the phenotype, which may be binary, continuous, or censored. Let $m_{ij} = (m_{ij1}, m_{ij2}, \ldots, m_{ijM})$ denote $M$ SNP markers ($M = 2{,}000$ for the Poly2000 GeneChip), where each $m_{ijk} = (m_{ijk1}, m_{ijk2})'$ represents the pair of alleles forming the marker genotype and each $m_{ijkl} = 1$ or $m_{ijkl} = 0$ denotes the presence or absence, respectively, of the designated marker allele. Let $x_{ij}$ denote a vector of covariates such as known candidate genes or known environmental factors, which may influence the penetrance function jointly with the putative disease gene. For simplicity, we use the notation $D_i$, $M_i$, and $X_i$ to denote data $(d_{i1}, \ldots, d_{in_i})$, $(m_{i1}, \ldots, m_{in_i})$, and $(x_{i1}, \ldots, x_{in_i})$, respectively, from the $i$th family. The goal of the analysis is to locate the putative disease gene. The putative disease genotype consists of paired alleles, denoted by $g_{ij1}$ and $g_{ij2}$. Given the parental origin of each allele, the genotype may take one of four possible genotypes—a/a, A/A, a/A, or A/a—in which the first allele in the pair is paternal and the other is maternal. For simplicity, let $g_{ij} = (g_{ij1}, g_{ij2})$ denote the genotype at the putative disease locus. For all marker loci, the parental origins of marker genotypes may be partially determined; for example, the parental origin of an allele in a homozygous genotype is known, and the parental origin of an allele in a heterozygous genotype in a nonfounder may be determined by comparing it with his or her father's or mother's marker genotypes. For the remaining markers, let $p_{ijk} = 0$ or $p_{ijk} = 1$ at the $k$th marker locus denote that the first allele is maternally derived or paternally derived, respectively (phase indicator). Let $p_{ij} = (p_{ij1}, \ldots, p_{ijM})$. Now, we let $G_i$ and $P_i$ denote $(g_{i1}, \ldots, g_{in_i})$ and $(p_{i1}, \ldots, p_{in_i})$, respectively.

### *An Estimating Equation Method*

The primary objective of linkage/linkage-disequilibrium analysis is to estimate RFs and ORs. Traditionally, the maximum-likelihood approach is used as a method

to estimate these parameters. Although conceptually straightforward, this approach has experienced many challenges, including violations to distributional assumptions and excessive computational burden. To overcome these problems, we have proposed an estimating-equation approach (Zhao et al. 1998, and in press). This approach is made possible by the score-estimating equation of the likelihood approach, without requiring all of the distributional assumptions, the conditional independence in particular. Hence, the estimating-equation approach is semiparametric and is expected to be more robust than the likelihood approach. In the Appendix, we have derived a score-estimating equation for estimation of the RFs and the ORs, and we have computed key quantities for a nuclear family. Specifically, the score-estimating equation for one marker locus may be written as

$$\varepsilon\left\{\begin{bmatrix}\sum_{j\in\mathcal{T}} Z_j' F_j \\ \sum_{j\notin\mathcal{T}} \frac{1}{\theta(1-\theta)}(Y_j - R_j\theta)\end{bmatrix}\,\middle|\, D,M,X\right\}\quad, \qquad (2)$$

where $\varepsilon(\cdot\,|\,D,M,X)$ is the conditional expectation of the "full score function," given observed data; $\Sigma_{j\in\mathcal{T}}$ and $\Sigma_{j\notin\mathcal{T}}$ are summations over founders and nonfounders, respectively; $F_j$ is a vector of residuals; $Z_j$ is the corresponding design matrix (to be detailed below); $R_j$ is the count of informative meioses on both paternal and maternal sides; and $Y_j$ is the count of recombinants from both sides. This simple score-estimating equation can now be used to construct the estimating equation for both two-point linkage analysis and then multipoint linkage analysis.

*Estimating Equation for Two-Point Linkage Analysis.*—As a relatively recent development in statistics, the estimating-equation technique has been successfully applied to a number of biological problems (Liang and Zeger 1986; Zhao et al. 1992). Conceptually, the estimating equation is defined as an equation that yields a consistent and normally distributed estimate. In the current context, one can take the above-described score function (2) as an estimating equation, without making the same distributional assumptions that are used to derive the score-estimating equation. The following discussion centers around the description of the estimating equation for two-point linkage analysis ($M = 1$).

Suppose that there is a family study with $I$ independent families and observed data $(D_i,M_i,X_i)$, as defined above. Following the construction of score function (2), one may choose the estimating equation for $(\alpha,\beta,\theta)$ to be

$$u(\alpha,\beta,\theta) = \sum_i u_i(\alpha,\beta,\theta)$$

$$= \sum_i \varepsilon^*\left\{\begin{bmatrix}\sum_{j\in\mathcal{T}_i} u_{ij}(\alpha,\beta) \\ \sum_{j\notin\mathcal{T}} u_{ij}(\theta)\end{bmatrix}\,\middle|\, D_i,M_i,X_i,\right\} = 0\ , \quad (3)$$

where the superscript "*" in $\epsilon^*(\cdot\,|\,D_i,M_i,X_i)$ is used to differentiate this "expectation" from the expectation used in the score function, where $u_{ij}(\alpha,\beta)$ and $u_{ij}(\theta)$ are estimating functions for estimation of $(\alpha,\beta)$ and $\theta$, respectively. These estimating functions may be written, respectively, as $u_{ij}(\alpha,\beta) = Z_{ij}'F_{ij}$ and $u_{ij}(\theta) = Y_{ij} - R_{ij}\theta$, where

$$F_{ij} = \begin{bmatrix} m_{ij1} - \mathrm{E}(m_{ij1}|\ p_{ij}\ ,g_{ij}) \\ m_{ij2} - \mathrm{E}(m_{ij2}|\ p_{ij}\ ,g_{ij})\end{bmatrix}$$

and

$$Z_{ij} = \begin{bmatrix} 1 & 1 \\ p_{ij}g_{ij1} + \bar{p}_{ij}g_{ij2} & p_{ij}g_{ij2} + \bar{p}_{ij}g_{ij1}\end{bmatrix},$$

$\mathrm{E}(m_{ijk}|p_{ij},g_{ij}) = P(m_{ijk} = 1|p_{ij},g_{ij})$ is specified by logistic regression (1) ($k = 1$ and $k = 2$). If the primary interest is in the map distance, $\theta = M(\delta)$, estimating equation (3) can still be used, after a modification to $u_{ij}(\theta)$, via $u_{ij}(\delta) = \{[\partial M(\delta)]/\partial\delta\}[Y_{ij} - R_{ij}M(\delta)]$, which may be thought of as a simple transformation from $\theta$ to $\delta$. Obviously, estimating equation (3) is fully specified, provided that the "expectation," $\epsilon^*(\cdot\,|\,D_i,M_i,X_i)$, is available. Hereafter, this "conditional expectation" is referred to as "E*."

The choice and computation of E* have been described elsewhere (Zhao et al. 1998, and in press). Fundamentally, the only requirement in choosing E* is that its marginal expectation, with respect to marker data, given the phenotypes, equals zero. For example, if the distribution $f(G_i,P_i|\,D_i,M_i,X_i)$ is specified, the conditional expectation in the score function can be chosen as E*. The only challenge to this choice is the computation, which involves summation over all putative disease genes and related phases and which potentially increases, at an exponential rate, with the number of founders. To avoid the above-described distributional assumption and to reduce the computational burden, we have introduced another approach to choosing and computing this expectation (Zhao et al. 1998, and in press). Our approach first decomposes a pedigree into a series of successive nuclear families (a nuclear family precedes another if one child in the former nuclear family is a parent in the latter nuclear family). Within each nuclear family, one computes the corresponding E* with a distribution func-

tion that is appropriate for that nuclear family. Consequently, the computational burden increases linearly with the number of nuclear families.

Estimating equation (3) has a simple and intuitive expression but may lead to the deceptive impression that recombination fractions and linkage-disequilibrium odds ratios are uncorrelated, because ORs are estimated by means of data from founders and RFs are estimated from data from nonfounders. This impression would be correct if all putative disease genotypes and parental origins of all marker alleles were observed. In general, however, estimated RFs and ORs are correlated: the ORs, if different from those of the null hypothesis, would help in the identification of haplotypes among founders and hence would improve the efficiency of the estimation of the RFs. On the other hand, the RFs, if close to zero, would help in the identification, on the basis of information about their children's haplotypes, of haplotypes among founders. These interdependence relationships are implicitly specified via expectation $\epsilon^* \mid D_i,M_i,X_i)$.

*Estimating Equations for Multipoint Linkage Analysis.*—The Poly2000 GeneChip provides densely distributed SNP markers, which are expected to provide more information via multipoint linkage than via two-point linkage analysis. As noted earlier, in the description of multipoint linkage analysis, the multipoint analysis to be considered here is to estimate an averaged RF, $\bar{\theta}_l$, at the $l$th locus, with its adjacent $c$ SNP loci. Now, let $\beta = (\beta_1, \ldots, \beta_c)'$ denote the log ORs between the putative disease gene and all $c$ neighboring SNP makers. Similarly, let $\alpha = (\alpha_1, \ldots, \alpha_c)'$ denote the corresponding intercepts in logistic regression (1).

Following a similar derivation for the two-point linkage analysis, we have derived the score-estimating equation under a likelihood function with multiple loci (Zhao et al., in press). Although the general formulation is complicated by the presence of the interference, it shares an expression similar to that of score function (2), when interferences between loci are absent. Intuitively, the estimating equation for the multipoint loci can be naturally viewed as an extension of estimating equation (3). Suppose that there is a family study with $I$ independent families and observed data $(D_i,M_i,X_i)$, as defined above. Extending estimating equation (3), one may estimate $(\alpha,\beta,\bar{\theta}_l)$ for the multipoint linkage analysis, via solution of the following estimating equation:

$$u(\alpha,\beta,\theta) = \sum_i u_i(\alpha,\beta,\theta)$$

$$= \sum_i \varepsilon^* \left\{ \begin{bmatrix} \sum_{j \in \mathcal{T}_i} Z'_{ij} F_{ij} \\ \sum_{j \notin \mathcal{T}} \sum_{k=1}^{c} (Y_{ijk} - R_{ijk}\bar{\theta}_l) \end{bmatrix} \middle| D_i,M_i,X_i \right\} = 0 , \qquad (4)$$

where $Z'_{ij}F_{ij}$ is a column vector of $Z_{ij1}F_{ij1}, \ldots,$ and $Z_{ijc}F_{ijc}$ and each $Z_{ijk}$ and $F_{ijk}$ are defined as in estimating equation (3), although for the $k$th marker locus. Note that the summation, $\sum_{k=1}^{c}$, is introduced as a derivative function of those RFs with respect to $\bar{\theta}_l$, which equals a vector of ones.

Because of the linearity in the expression of $Z'_{ij}F_{ij}$ and $\sum_{k=1}^{c}(Y_{ijk} - R_{ijk}\bar{\theta}_l)$, one can move the expectation so that it is locus specific, resulting in

$$u(\alpha,\beta,\theta) = \sum_i \left\{ \begin{bmatrix} \sum_{j \in \mathcal{T}_i} \varepsilon_1^*(Z'_{ij1}F_{ij1} \mid D_i,M_i,X_i) \\ \varepsilon_c^*(Z'_{ijc}F_{ijc} \mid D_i,M_i,X_i) \end{bmatrix} \right\}$$
$$\left\{ \sum_{j \notin \mathcal{T}} \sum_{k=1}^{c} \varepsilon_k^*(Y_{ijk} - R_{ijk}\bar{\theta}_l \mid D_i,M_i,X_i) \right\}$$

in which all E* values can be computed as specific to each locus. This formulation suggests that one can now compute a locus-specific E*, $\varepsilon_k^*(\cdot \mid D_i,M_i,X_i)$ for each individual marker locus, rather than having to sum over all possible configurations of $c$ SNP marker loci. Hence, the computational burden of the multipoint linkage analysis increases with the number of loci, $c$. The locus-specific calculation of these expectations, $\varepsilon_k^*(\cdot \mid D_i,M_i,X_i)$, follows that in the two-point linkage analysis.

### A Mapping Strategy

The preceding section has described an estimating equation that can be used to estimate RFs and ORs for the linkage and linkage-disequilibrium analyses jointly. The estimating-equation approach is applicable to all types of pedigrees with various phenotypes and allows one to incorporate candidate genes and environmental factors, and its computational burden increases linearly with the sizes of the families and with the number of loci in the analysis. Using this approach, one can design various mapping strategies with estimated RFs and ORs. The strategy to be considered here is to jointly estimate RFs and ORs throughout the genome and to estimate averaged RFs on the basis of several adjacent SNP loci.

Because of the joint estimation of RFs and ORs, both statistics can be used for making inferences about signals arising from the presence of linkage, from the presence of linkage disequilibrium, or from both. Figure 2 identifies four regions indexed by estimated RFs and ORs; when both the RF and the log OR are at approximately their respective null values, there is apparently no signal. On the other hand, the presence of both linkage and linkage disequilibrium yields a strong signal. In practice, this signal could be captured either by linkage only or by linkage disequilibrium only. In figure 2, a fictitious estimator, with its confidence region plotted, shows presence of the signal. However, if only the signal captured by linkage or only the signal captured by linkage disequilibrium signal is considered, it may not be detectable at a statistically significance level. Note that use of joint

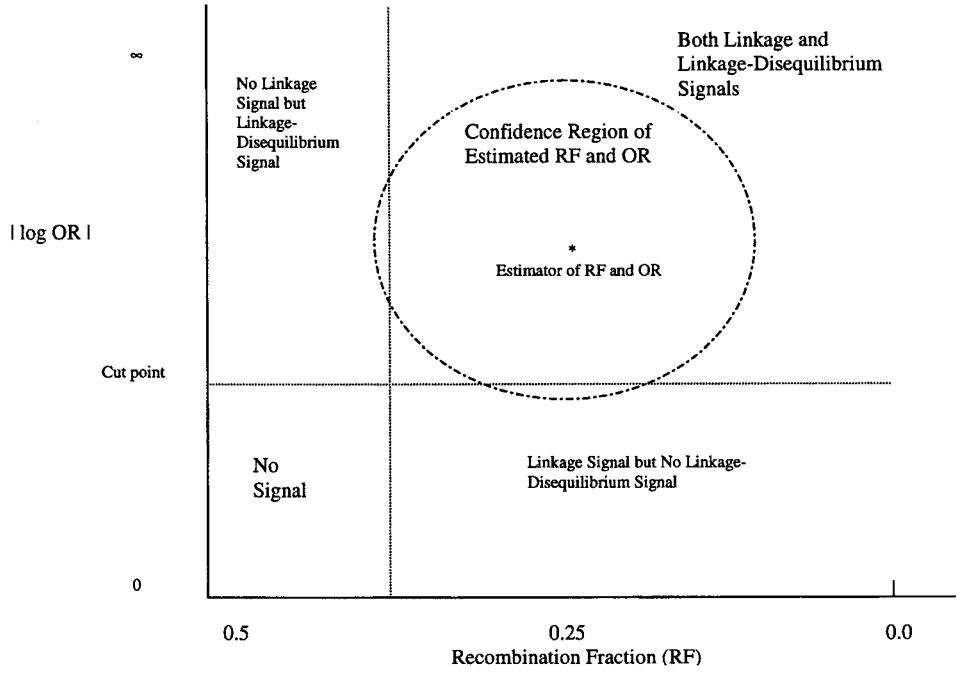


**Figure 2** Illustration of a combined linkage/linkage-disequilibrium analysis with one SNP marker

estimation and inference has a motivation similar to that underlying the disequilibrium-transmission test (TDT) (Self et al. 1991; Ewens and Spielman 1995; Risch and Merikangas 1996). The choice of this mapping strategy dictates both how to estimate $(\hat{\alpha},\hat{\beta},\hat{\theta})$, and how to make an inference about the putative disease genes.

### *Estimation and Inference*

Following the mapping strategy outlined above, this section describes both an algorithm for estimation of parameters $(\hat{\alpha},\hat{\beta},\hat{\theta})$ and the procedure for making an inference. Just as in the maximum-likelihood calculation, the estimator $(\hat{\alpha},\hat{\beta},\hat{\theta})$ satisfies the estimating equation, $u(\hat{\alpha},\hat{\beta},\hat{\theta}) = 0$, and thus is a solution to this equation. Since there is no explicit solution to this equation, one generally has to solve the equation iteratively, using an algorithm such as the Newton-Raphson method. After experiencing numerical instability with a naive joint estimation of $(\hat{\alpha},\hat{\beta},\hat{\theta})$, we have modified the Newton-Raphson algorithm by enumerating a series of RF values and iteratively estimating $(\alpha,\beta)$. Consider an estimation with $c$ adjacent loci, estimating an averaged RF jointly with vectors $\alpha = (\alpha_1,\alpha_2,\ldots,\alpha_c)'$ and $\beta = (\beta_1,\beta_2,\ldots,\beta_c)'$. Profiling over a series of RFs between 0 and .5—for example, $\theta_0$—one solves the estimating equation corresponding to $(\alpha,\beta)$. Starting from $(\alpha_0,\beta_0)$, one iterates to a new value, $(\alpha_1,\beta_1)$, via

$$\begin{pmatrix}\alpha_1\\ \beta_1\end{pmatrix} = \begin{pmatrix}\alpha_0\\ \beta_0\end{pmatrix} - \left[\frac{\partial u(\alpha_0,\beta_0|\theta_0)}{\partial(\alpha_0,\beta_0)}\right]^{-1} u(\alpha_0,\beta_0|\theta_0)\ ,$$

where $u(\alpha_0,\beta_0|\theta_0)$ is the part of estimating equation (4) corresponding to $(\alpha,\beta)$ and where all quantities on the left-hand side of the equation are evaluated at the initial value $(\alpha_0,\beta_0)$. This iterative procedure continues until the convergence of all elements in $(\alpha,\beta)$; at the convergence, the estimated parameter is a function of $\theta_0$, denoted by $\alpha(\theta_0)$ and $\beta(\theta_0)$. With this estimate, one can now evaluate the estimating function $u(\theta_0|\alpha_0,\beta_0)$, which is part of estimating equation (4) corresponding to $\theta$. By directly inspecting this estimating function $u(\theta_0|\alpha_0,\beta_0)$ over a range of $\theta_0$, one can immediately identify the solution at which $u(\hat{\theta}|\hat{\alpha},\hat{\beta}) = 0$.

Estimated parameters have desirable asymptotic properties that can be used for making inference. As noted above, the estimating function results from the summation of independent pedigrees. As the number of independent pedigrees becomes large, the estimator has an asymptotic-normal distribution, following directly from the central-limit theory applied to the summation of independent pedigree-specific estimating functions, $u_i(\alpha,\beta,\theta)$. As long as $E[u_i(\alpha,\beta,\theta)] = 0$, the estimating function, $\sum_i u_i(\alpha,\beta,\theta)$, has an asymptotic-normal distribution with zero mean and with an asymptotic-variance matrix that can be easily estimated by $I^{-1}\Sigma_i u_i(\hat{\alpha},\hat{\beta},\hat{\theta})u_i(\hat{\alpha},\hat{\beta},\hat{\theta})'$.

This asymptotic-normal distribution can now be used to construct a test statistic that is comparable to the score test (e.g., see Zhao et al. 1998, and in press). By Taylor expansion, one can show that the estimator is consistent and has an asymptotic-normal distribution, by means of the asymptotic-variance matrix

$$I\left[\sum_i \frac{\partial u_i(\hat{\alpha},\hat{\beta},\hat{\theta})}{\partial(\alpha,\beta,\theta)}\right]\left[\sum_i u_i(\hat{\alpha},\hat{\beta},\hat{\theta})u_i(\hat{\alpha},\hat{\beta},\hat{\theta})'\right] \times\left[\sum_i \frac{\partial u_i(\hat{\alpha},\hat{\beta},\hat{\theta})}{\partial(\alpha,\beta,\theta)}\right].$$

This asymptotic-normal distribution can be used to construct test statistics for the estimator, as well as for the confidence region, such as those shown in figure 2.

It is important to recognize, however, that some linkage studies have a limited number of families but include multigenerational families with many founders. Although the tendency has not been rigorously proved, conventional wisdom suggests that the estimated parameters should approach an asymptotic-normal distribution as family size—and, hence, the number of founders—becomes large (R. Elston, personal communication). Nevertheless, this claim for the asymptotic property remains to be verified.

### *Simulation Study*

We consider two limited simulation studies, (*a*) to assess asymptotic properties within the finite sample and (*b*) to illustrate the utility of the technique for the mapping of complex traits by means of SNP markers. Throughout both simulation studies, we consider 200 nuclear families, each with two parents and four children. Families are ascertained only if they include three or more affected members. In assessing finite-sample properties, the first study simulates a map of 10 SNP loci, in which an equal distance of 10 cM is assumed. Using this map, we suppose that this putative disease gene conveys a binary phenotype with log OR = 5; for example, carriers confer 73% risk, and noncarriers confer 2% risk. To differentiate between this OR in the penetrance and the log OR $\beta$ for linkage disequilibrium, let us refer to this as the "log penetrance OR," which hereafter will be denoted as " $\lambda$." The allele frequency is assumed to be .01. The simulated putative disease-gene locus is between the second and third loci, with RF = 6 cM, and the putative disease allele is in linkage disequilibrium with the marker alleles in adjacent loci and has a log OR of $\beta = 2$. The study has 50 replicates. In each replicate, the study simulates nuclear-family data and estimates both RFs and ORs. In this study, we chose to pool two adjacent loci and to estimate pooled RFs. These estimates are used to evaluate finite-sample properties of interest, via evaluation of their distributions (fig. 3).

To illustrate this new method for the mapping of complex traits by means of SNP markers, the second study simulates a map of 44 SNP loci with an equal distance of 5 cM, representing an average number of Poly2000 GeneChip SNP loci per chromosomal arm (shown in fig. 4). This study simulates one genome-scan experiment, just as if a single linkage study were being conducted. In this example, we simulate two major genes, one of which is between locus 4 and locus 5, with $\theta = .03$ on both sides, and the other of which is between locus 24 and locus 25, with $\theta$ values of .01 and .05, respectively. The first putative disease gene has a modest penetrance, with $\lambda = 4$ among carriers, and allele frequency .01. The second putative disease gene confers a low penetrance, with $\lambda = 2.3$ among carriers, and allele frequency .05. Both genes are in linkage disequilibrium with their adjacent SNP markers, with log OR of $\beta = 2$. To improve the efficiency of the estimation, we estimate the averaged RF with an adjacent-locus SNP marker. Because there are two major genes, any linkage analysis of one putative disease gene would consider the other gene as the presence of the genetic heterogeneity. To retain the realistic aspect of linkage analysis, we perform two linkage analyses under their respective segregation models—that is, a single-gene model with $\lambda = 4$ assumed and allele frequency .01 and another single-gene model with $\lambda = 2.3$ and the allele frequency .05. Consequently, for the analysis, both single-gene models are considered to be an incorrect specification of two-gene models. Additional genome-scan experiments on other complex traits have also been conducted and analyzed, and results from them will be reported elsewhere (Zhao et al., in press).

## Results

### *Finite-Sample Properties*

Figure 3 shows distributions of estimated averaged RFs for paired SNP loci—that is, loci 1 and 2, loci 2 and 3, etc.—in the two left-hand columns, as well as estimated ORs for each SNP locus, in the two right-hand two columns. To examine the distributions of these estimates, we compute the density functions of these limited replicates, using kernel estimation (e.g., see Silverman 1990). By observing RF distributions, one can see that the distributions of estimated ORs are approximately truncated normal distributions. The distribution of the averaged RF with loci 2 and 3 is on the far left, and those with loci 1 and 2 and loci 3 and 4 also show

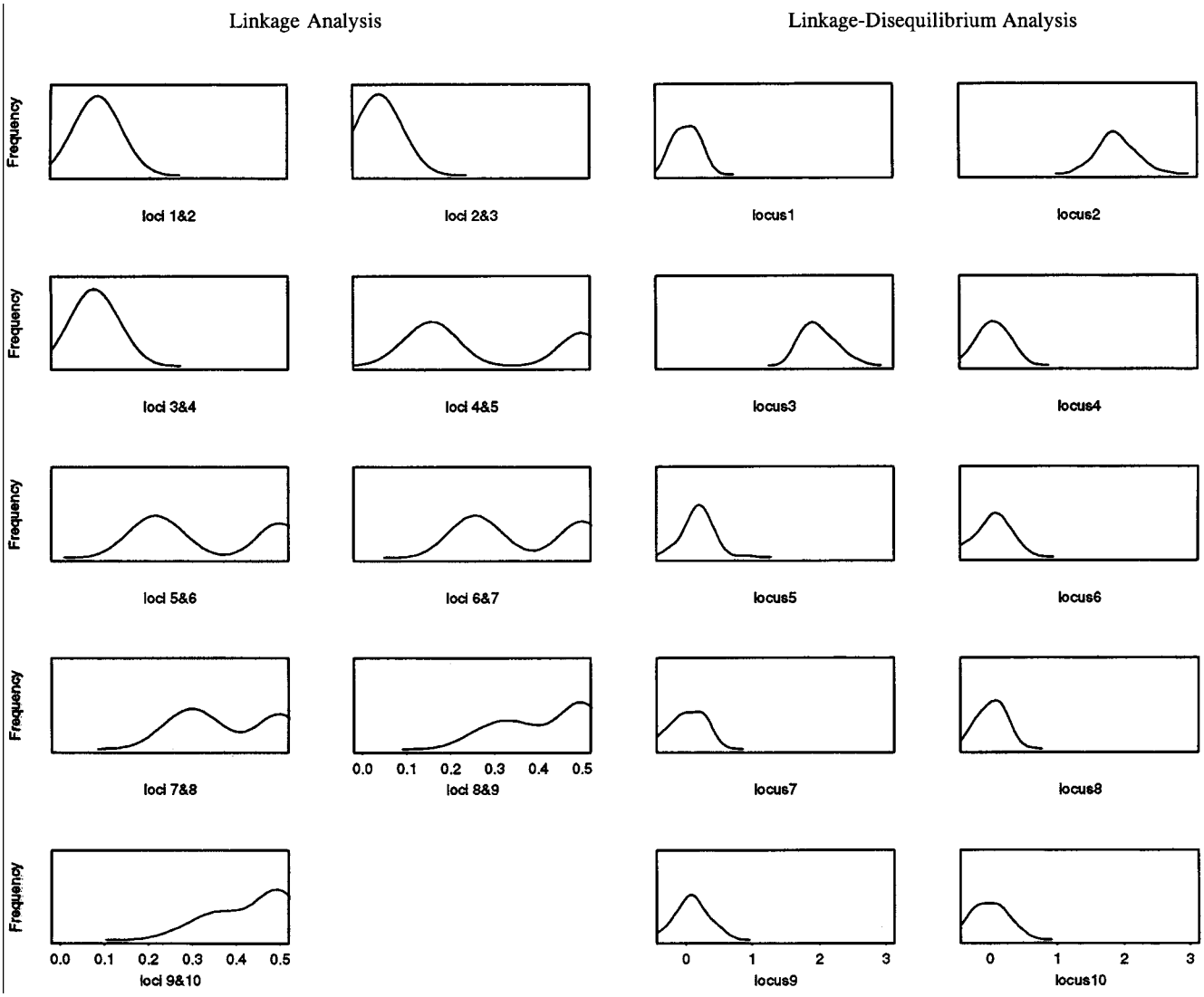


**Figure 3** Distributions of estimated RFs (*left-hand two columns*) from linkage analysis and of estimated log ORs (*right-hand two columns*) from linkage-disequilibrium analysis, estimated on the basis of 50 replicates in the simulation study. In the left-hand two columns, the pattern of these distributions indicates that the putative disease locus is closer to marker loci 2 and 3 and is farther from both marker locus 1, on the left-hand side, and marker loci 4–10; in the right-hand two columns, the pattern of these distributions indicates that the putative disease allele is in linkage disequilibrium with alleles at marker loci 2 and 3 and is in linkage equilibrium with alleles at the other marker loci.

positive linkage signals. The distributions of the averaged RFs for loci 4 and 5, loci 5 and 6, loci 6 and 7, loci 7 and 8, loci 8 and 9, and loci 9 and 10 shift toward the null, indicating that these loci are increasingly distant from the putative disease locus. However, their bimodality causes some concern, since this phenomenon needs special attention when confidence bands for linkage-based genome scans are estimated.

Distributions of the estimated ORs offer much clearer signals with respect to the position of the putative disease gene. Distributions of the ORs for loci 2 and 3 center around 2, the true value, and appear to be symmetrical. On the other hand, distributions of ORs for loci 1 and 4–10 center around 0, indicating linkage equilibrium between the putative disease allele and all these SNP loci.

### *A Simulated Genome-Scan Experiment*

Figure 5 shows estimated RFs and ORs for 44 SNP loci; the dotted and dashed lines are for estimates obtained under the segregation models for the modest-penetrance (first) and the low-penetrance (second) disease genes, respectively. The top panel of figure 5 shows the estimates of RF. It appears that the estimates under the

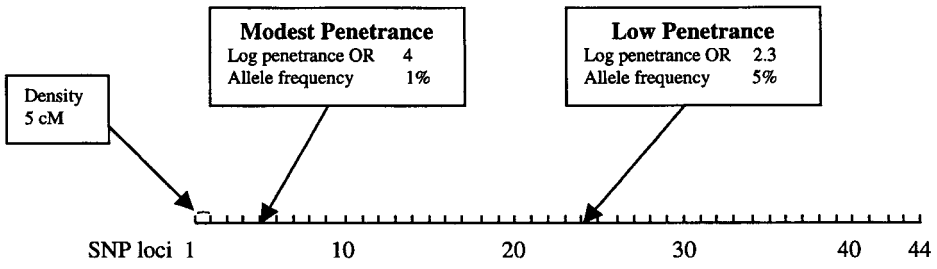


**Figure 4** Simulation experiment generating a map with 44 SNP markers and localizing two major disease genes, the first of which has a modest penetrance and is between marker loci 4 and 5 and the second of which has a low penetrance and is between loci 24 and 25. The objective of this simulation experiment is to illustrate the utility of the semiparametric method in the mapping of a complex trait via a genomewide linkage/linkage-disequilibrium analysis.

modest-penetrance model are biased toward null. However, the linkage signals, especially those for the first putative disease gene, are rather noticeable. Throughout the genome, these estimates are more biased toward the null than are those under the with low-penetrance segregation model. However, under the second model, it appears that there are more false-positive leads. Nevertheless, it is important to recognize that linkage analysis has not missed the signals of both putative disease genes, despite the misspecification of the segregation models.

The middle panel of figure 5 shows the estimated log ORs under the two segregation models. In a reversal of the pattern of biases, estimated log ORs under the second model tend to be more biased toward the null than are those under the first model. Nevertheless, both sets of estimates show a consistent pattern of linkage-disequilibrium signals and have correctly detected two putative disease loci, with a couple of false-positive leads. At the detected loci, estimated log ORs are ~0.5, which is substantially biased away from the true value of 2. This observation indicates that misspecification of the penetrance function has substantial influence on the estimation of ORs.

Multiplying the log OR and (.5 − RF) results in combined linkage/linkage-disequilibrium estimators, which are shown in the bottom panel of figure 5. Interestingly, this combined estimator detects precisely the locations of both putative disease genes. Moreover, it appears that estimators under two separate segregation models give fairly comparable results, demonstrating the robustness property against the misspecification of the segregation model. Although favorable, this result should be interpreted in the context of this particular simulation experiment and should not be overly interpreted at this time (see the Discussion section, below).

## Discussion

This article has introduced a method for the mapping of complex traits via linkage analysis and linkage-disequilibrium analysis, with SNP markers. In comparison with other methods for the mapping of complex traits, this approach takes advantage of the biallelic nature of SNP markers, so that linkage disequilibrium can be naturally quantified via ORs, and the computational burden therefore is reduced. The joint estimation of the linkage/linkage-disequilibrium parameters allows one to make inferences about both aspects of the complex traits simultaneously, yielding more information. Furthermore, taking advantage of the high density of the SNP markers, one can pool several adjacent SNP loci in estimation of an averaged RF, thereby improving efficiency. Since this approach is semiparametric, it requires weaker assumptions than are required by the likelihood method, and it is more efficient than model-free (or nonparametric) methods. Besides, as a special member in this semiparametric framework, it retains all of the desirable properties. Specifically, this approach can handle binary, continuous, and censored phenotypes; allows one to incorporate all covariates in the investigation of gene/gene and gene/environmental interactions; and is able to take into account any pedigree structures (e.g., relative pairs, affected-relative clusters, nuclear families, or extended pedigrees). A limited simulation study indicates that the estimator has desirable finite-sample properties. Analyzing the simulated example, we have shown that (*a*) linkage analysis has correctly detected the presence of both putative disease loci, along with a few false-positive signals, (*b*) linkage disequilibrium seems to suggest the correct locations of both putative disease genes, and (*c*) the combined linkage/linkage-disequilibrium analysis has given a clear signal for both putative disease genes, in the presence of the genetic heterogeneity. This result suggests that SNP markers, if analyzed appropriately, can be used effectively for the mapping of complex traits.

We need to interpret these results cautiously, recognizing the limitations of our simulation experiments. Let us specifically consider alternative scenarios, in which the proposed approach may fail. First, if the putative

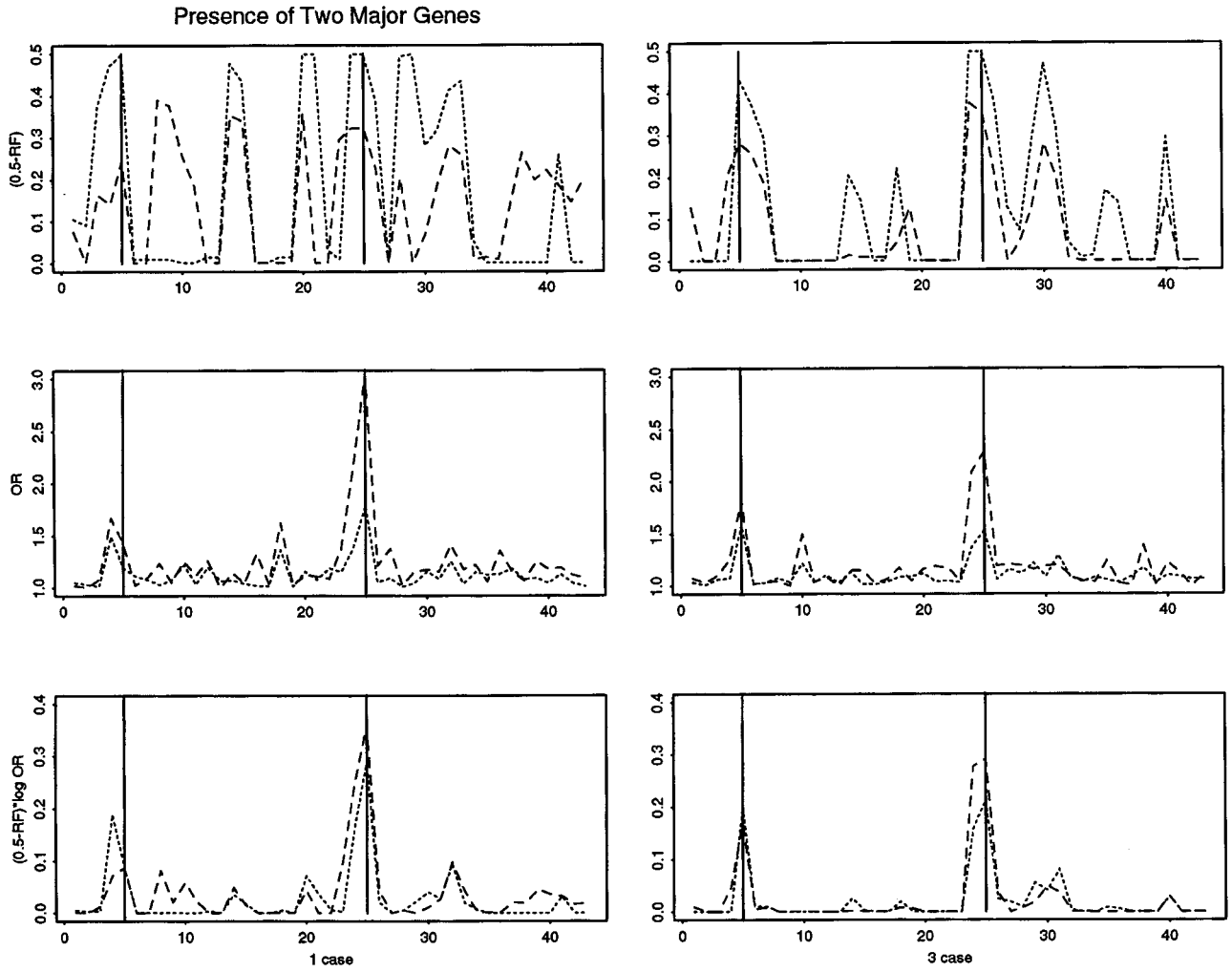


**Figure 5** Results obtained from a genome-scan experiment based on linkage analysis, linkage-disequilibrium analysis, and combined linkage/linkage-disequilibrium analysis. The dotted lines indicate the results obtained with the segregation model for the gene with modest penetrance (between loci 4 and 5), where the dashed lines indicate those obtained with the segregation model for the gene with low penetrance (between loci 24 and 25).

disease gene is in linkage equilibrium with adjacent SNP markers, the linkage-disequilibrium test, as well as the combined linkage/linkage-disequilibrium test, would inevitably fail to detect any relevant signals. Hence, the only reliable information is the linkage results, which may include many false-positive findings, as has been shown in the simulation experiment. It is also known, from population-genetics theory, that linkage equilibrium may well occur if the mutations have been introduced into the population for "millions of generations" or if multiple mutations have occurred during evolution. Second, the marker frequency at all SNP loci is assumed to be .50, which is optimal for linkage analysis. In practice, only an extreme deviation from the .50 frequency may decrease the efficiency of linkage analysis—and, hence, may fail to detect linkage signals. Third, the simulation study has a full knowledge of the true genetic mechanism and has specified two segregation models that approximate the underlying true models. Without this information, an incorrect segregation model may mislead the analysis; of course, other model-based methods for the mapping of complex traits face the same difficulty. Fourth, the genome-scan experiment simulates nuclear-family data, which are most favorable for linkage-disequilibrium analysis but not for linkage analysis. If large but few families are included, such simulated data may have more information about linkage than about linkage disequilibrium. If linkage and linkage disequilibrium are naively combined, one may fail to localize putative disease loci.

In choosing a measurement for linkage disequilibrium, we have considered several alternative measurements be-

sides log OR $\beta$. Traditionally, the correlation coefficient is often used to characterize an *empirical association* between two random variables. However, the value of the correlation coefficient is generally bounded by the marginal frequencies. In the case of SNPs, allele frequencies vary widely among SNP loci, and thus the correlation coefficients for SNP alleles and the putative disease allele have different boundaries, which prohibits a direct comparison among correlations that have different SNPs.

Besides the correlation coefficient and OR, several other measures have been introduced to quantify linkage-disequilibrium in the context of population genetics; they have been reviewed by Devlin and Risch (1995) and have then been further studied by Guo (1997). Although valuable, their conclusions from both analytic and simulation studies should be interpreted in the context of the specific evolutionary model. For example, the covariance between marker and putative genotype has been shown to be preferred in the specific evolutionary model (Devlin and Risch 1995). However, the covariance as an empirical measurement of an association between two variables is known to be sensitive to the marginal means of these variables, and, furthermore, if two variables are binary, the covariance is further bounded by functions of these means. These restrictions would increase difficulties in the comparison of covariances from many SNP loci that have different allele frequencies.

In designing our simulation study, we have chosen the study design with nuclear families with multiple affects, primarily for the simplicity and efficiency in the completion of the calculation. Because the semiparametric framework is robust against the ascertainment bias and because the nuclear family is the fundamental structure for all pedigree analysis, we conjecture that observations made in this paper are likely to hold. Nevertheless, it is important to assess the utilities of SNP markers and of this semiparametric method under different design scenarios—for example, affected sibling pairs, nuclear families with a single ascertainment, extended pedigrees, and mixtures of these family structures. This work will be performed and reported in the future.

This semiparametric method is closely related to methods that have been developed, in the past, for linkage analysis and for linkage-disequilibrium analysis. With regard to linkage/linkage-disequilibrium analysis, earlier methods include those model-free methods that do not require specification of the mode of inheritance. For example, in linkage studies of affected relative pairs, one method tests whether the observed marker loci in the relative pairs are more identical by descent than by chance (Kruglyak et al. 1995; Kruglyak and Lander 1995; Whittemore 1996). Another method is the transmission/disequilibrium test (TDT), which is designed to test the presence of both linkage and linkage disequilibrium between a marker locus and putative disease locus (Self et al. 1991; Spielman et al. 1993; Schaid and Sommer 1994; Ewens and Spielman 1995; Risch and Merikangas 1996; Martin et al. 1997). Although robust, model-free methods suffer potential loss of efficiency because they do not account for any knowledge about disease etiology. Alternatively, either model-based (or parametric) methods based on likelihood theory or LOD-score methods can also be used for the mapping of complex traits, but they require assumptions of segregation models for putative disease genes. Model-based linkage analysis estimates genetic distances, or RFs, of genetic marker loci by means of a putative disease locus (Ott 1989). Similarly, model-based linkage-disequilibrium analysis estimates association parameters that quantify the deviation from linkage equilibrium. Such associations may result from cosegregation during evolution (Xiong and Guo 1997). Although efficient, model-based methods have (*a*) high computational needs and (*b*) strong distributional assumptions—in particular, the conditional independence of phenotypes within families, given putative disease genes for the likelihood methods.

Now let us compare the semiparametric method with these model-based and model-free methods, for mapping studies. First, compared with methods for linkage analysis, the semiparametric-method approach may be preferred to LOD-score methods, because of computational efficiency and inferential robustness, but may be criticized for the possible efficiency loss; compared with model-free methods, advantages of this approach include (*a*) efficiency gain, (*b*) ability to estimate RFs, and (*c*) ability to include putative disease genes that interact with covariates, including candidate genes and environmental factors. The main disadvantage is the requirement of assuming a major-gene model, which is not needed by model-free methods, partly because model-free methods test strictly under the null hypothesis (i.e., no linkage/linkage equilibrium). However, limited simulation studies that we have conducted indicate that this approach retains an asymptotic distribution under the null hypothesis, regardless whether segregation models are correctly specified.

Second, we compare the semiparametric method with methods for linkage-disequilibrium analysis. Population geneticists have developed several models, most of which have been reviewed by Xiong and Guo (1997). The primary advantage of population-genetic models is their efficiency in the mapping of complex traits, which becomes critical in fine-scale mapping. However, their approach requires several key assumptions about population genetics, such as the age of an isolated population and the new mutation rate in the study population. Nevertheless, when these assumptions are met—for example,

in studies on "young" and "isolated" populations (e.g., Finnish communities)—their approach becomes extremely efficient, because of incorporation of the evolutionary history of the population. In studies of outbred population, such as the U.S. populations, their approach is still applicable, but the results must be interpreted cautiously, since several key assumptions may be violated. In contrast, the method described here captures an *empirical association* between putative disease alleles and the SNP marker alleles, which results from population-genetic forces, population admixture, or other factors. Hence, this approach is expected to be more robust than population-genetic models, and the robustness permits its application to studies of outbred populations. The primary trade-off is a potential loss in efficiency.

Third, this approach is conceptually equivalent to the TDT, since both tests can be used to detect combined linkage and linkage-disequilibrium signals. Their key difference lies in the fact that the TDT detects nonrandom transmission of disease alleles from founders to nonfounders, whereas the new method jointly estimates parameters for linkage and linkage disequilibrium, using both founders' and nonfounders' data, respectively. The primary advantage of the TDT is its robustness, since it requires no assumptions about the putative disease genes. However, the TDT does not utilize available phenotype data from founders and hence is less efficient than the new approach. Furthermore, the TDT may fail to locate the disease genes by means of markers that are closely linked but are in linkage equilibrium.

The preceding discussion clearly indicates similarities and differences between this semiparametric method and many established model-based and model-free methods. To gain further insights, it is essential to undertake numerical comparison among these different approaches, under different sampling scenarios. This activity will be undertaken in the future.

In conclusion, the localization of putative disease genes, without knowledge of their penetrances and corresponding allele frequencies, is challenging and requires multiple strategies to detect their signals. SNP-chip technology promises a high throughput and efficient and complete genotyping. With chip-based SNP markers, this new method can be effectively used to localize putative disease genes, by means of both linkage signals and linkage-disequilibrium signals. The limited simulation experiment indicates the feasibility of the mapping of complex traits such as cancer and coronary heart disease.

## Acknowledgments

The authors would like to thank Dr. David Wang for helpful discussions regarding SNP markers and chip technology. This research is supported in part by National Institutes of Health grants CA55670, CA64046, CA-33619, CA-53996, AG15358, and AG14358.

## Appendix

### A Likelihood and Its Score Function for Randomly Ascertained Families

Consider a likelihood of the observed data $(D,M,X)$ for one pedigree with one SNP marker,

$$f(D,M,X) = \sum_{G,P} f(D,M,X,G,P) \propto \sum_{G,P} f(D|X,G)f(M,G,P) \ ,$$

where the first equality holds for the marginal distribution and where the second proportionality holds, under the assumption that phenotypes depend only on putative disease genes and covariates and that covariates are independent of markers and phase indicators. The joint distribution $f(D\,|\,X,G)$ of the phenotype, given the putative disease gene and covariates, is obtained as a product of marginal probabilities, under the assumption of conditional independence; that is, $f(D|X,G) = \prod_j f(d_j|x_j,g_j)$.

The joint distribution $f(M,G,P)$ is indexed by both the RF and the OR and is obtained via a decomposition based on the pedigree (in this context, a pedigree is defined as a completely connected family tree). Every pedigree has founders—those individuals whose parents are not included in the pedigree—and nonfounders. Let $\mathcal{F} = (j \mid j$ is founder in the pedigree) represent a set of founders. On the basis of genetic transmission, the joint distribution $f(M,G,P)$ may be decomposed as

$$f(M,G,P) = \prod_{j\in\mathcal{F}} f(m_j|g_j,p_j)f(g_j)f(p_j) \prod_{j\notin\mathcal{F}} f(m_j,g_j,p_j|m_{[j]},g_{[j]},p_{[j]}) \ ,$$

where the notation $[j]$ denotes the subscripts of parents of the $j$th nonfounder. The distribution $f(g_j)$ among founders may be simplified to be the product $f(g_{j1})f(g_{j2})$, where $f(g_{j1})$ and $f(g_{j2})$ are binomial and where $f(p_j)$ has a uniform

distribution at the marker locus for the founder. So this distribution function is fully specified, given $f(m_j|g_j,p_j)$ for founders and $f(m_j,g_j,\ p_j|m_{[j]},g_{[j]},p_{[j]}\ )$ for nonfounders.

Among founders, the distribution $f(m_j|g_j,p_j)$ is specified, given the phase indicator ($p_j$) and the putative genotypes. If the phase indicator $p_j = 1$, then the paternal and maternal haplotypes are $g_{j1}m_{j1}$ and $g_{j2}m_{j2}$, respectively; otherwise ($p_j = 0$), and then the resulting haplotypes are $g_{j1}m_{j2}$ and $g_{j2}m_{j1}$, respectively. Hence, the distribution $f(m_j|g_j,p_j)$ may be represented as

$$f(m_j|g_j,p_j) = [f(m_{j1}|g_{j1})f(m_{j2}|g_{j2})]^{p_j}[f(m_{j1}|g_{j2})f(m_{j2}|g_{j2})]^{\bar{p}_j}\ ,$$

where $\bar{p}_j = 1 - p_j$, and the conditional distribution of the marker allele, given the putative disease allele—for example, $f(m_{j1}|g_{j1})$—can be modeled via logistic regression (1), which estimates log OR for linkage disequilibrium. After the logistic model is substituted into the above distribution, some algebraic simplification leads to

$$f(m_j|g_j,p_j) = \frac{\exp(\alpha m_j^* + \beta z_j)}{[1 + \exp(\alpha + \beta g_{j1})][1 + \exp(\alpha + \beta g_{j2})]}\ ,$$

where $m_j^* = m_{j1} + m_{j2}$, and $z_j = \ p_j\ (m_{j1}g_{j1} + m_{j2}g_{j2}) + \ \bar{p}_j\ (m_{j2}g_{j2} + m_{j2}g_{j1})$; and $g_{jl}$ here is coded to be 1 and 0, for the A and a alleles, respectively.

Among nonfounders, the joint distribution $f(m_j,g_j,p_j|m_{[j]},g_{[j]},\ p_{[j]}\ )$ is specified by the recombination process, with phase indicators and putative genotypes. Let $r_{jl} = 1$ and $r_{jl} = 0$ indicate, respectively, whether the paternal (maternal) meiosis is informative or noninformative. Given the nature of the SNP marker, $r_{jl} = 1$ only if both the SNP marker genotype and the putative genotype are heterozygous—that is, 0/1 and a/A. Resulting from this pair of SNP genotypes and putative genotypes are four possible haplotypes: 0a,1a,0A and 1A. Furthermore, let $y_{jl} = 1$ and $y_{jl} = 0$ indicate, respectively, that the informative paternal (maternal) meiosis is recombinant or nonrecombinant. By directly comparing a child's haplotype with his or her parents' haplotypes, one can determine the recombination status; $y_{j1} = 1$ if the paternal haplotype of the child is not the same as one of father's two haplotypes, and $y_{j1} = 0$ otherwise. Similarly, one can determine the maternal indicator $y_{j2}$. Under the assumption that paternal and maternal RFs are the same, the joint distribution $f(m_j,g_j,p_j|m_{[j]},g_{[j]},p_{[j]})$ may be represented as

$$f(m_j,g_j,p_j|m_{[j]},g_{[j]},p_{[j]}) = [\theta/(1-\theta)]^{Y_j}(1-\theta)^{R_j}\ ,$$

where $Y_j = r_{j1}y_{j1} + r_{j2}y_{j2}$ is the count of the recombination events and $R_j = r_{j1} + r_{j2}$ is the count of the informative meioses.

Taking the derivative of the log-likelihood function with respect to unknown parameters $(\theta,\alpha,\beta)$ results in the score function of interest, $\partial \log f(D,M,X)/\partial(\theta,\alpha,\beta)$, which may be written as

$$\sum_{G,P} f(G,P|D,M,X)\left[\sum_{j\in\mathcal{T}} \frac{\partial \log f(m_j|g_j,p_j)}{\partial(\theta,\alpha,\beta)} + \sum_{j\notin\mathcal{T}} \frac{\partial \log f(m_j,g_j,p_j|m_{[j]},g_{[j]},p_{[j]})}{\partial(\theta,\alpha,\beta)}\right],$$

where the first summation, $\Sigma_{G,P}$, is over all possible putative genotypes and phase indicators, the first interior summation, $\Sigma_{j\in\mathcal{T}}$, is over all founders, and the second interior summation, $\Sigma_{j\notin\mathcal{T}}$, is over all nonfounders. Since $\log f(m_j|g_j,p_j)$ is not a function of the RF value, its derivative with respect to $\theta$ equals zero, and an individual derivative with respect to $\alpha$ may be obtained as

$$\frac{\partial \log f(m_j,g_j,p_j)}{\partial\alpha} = [m_{j1} - \mathrm{E}(m_{j1}|p_j,g_j)] + [m_{j2} - \mathrm{E}(m_{j2}|p_j,g_j)] = (1,1)F_j\ ,$$

where $F_j = [m_{j1} - \mathrm{E}(m_{j1}|p_j,g_j),m_{j2} - \mathrm{E}(m_{j2}|p_j,g_j)]'$ is a vector of residuals, and

$$\mathrm{E}(m_{j1}|p_j,g_j) = P(m_{j1} = 1|g_{j1})p_j + P(m_{j1} = 1|g_{j2})\bar{p}_j\ ,$$

$$\mathrm{E}(m_{j2}|p_j,g_j) = P(m_{j2} = 1|g_{j2})p_j + P(m_{j2} = 1|g_{j1})\bar{p}_j\ ,$$

and the probability function is specified via logistic regression (1). The expression of the above-described score

contribution is consistent with the usual expression of the score function under the exponential family (Zhao and Prentice 1990). Now the derivative with respect to $\beta$ has a similar expression and may be written as $[\partial \log f(m_j,g_j,p_j)/\partial\beta] = (g_{j1}^*,g_{j2}^*)F_j$, where $g_{j1}^* = p_j g_{j1} + \bar{p}_j g_{j2}$ and $g_{j2}^* = p_j g_{j2} + \bar{p}_j g_{j1}$. Now let $Z_j$ denote a $2 \times 2$ matrix in which the first row is the vector (1,1) and the second row is the vector $(g_{j1}^*,g_{j2}^*)$. The above-described score function with respect to $(\alpha,\beta)$ can now be expressed in a familiar form, as $\partial \log f(m_j,g_j,p_j)/\partial(\alpha,\beta) = Z_j' F_j$.

Consider the second part of the above-described derivative of $\log f(m_j,g_j,p_j \mid m_{[j]},g_{[j]},p_{[j]})$. Because this joint distribution is not indexed by parameters $(\alpha,\beta)$, the corresponding derivative equals zero. The derivative with respect to $\theta$ can be obtained for the binomial distribution and may be expressed as $\{1/[\theta(1-\theta)]\}(Y_j - R_j\theta)$.

Integration of the two parts of the above-described derivative leads to a general expression of the score function,

$$\frac{\partial \log f(D,M,X)}{\partial(\theta,\alpha,\beta)} = \varepsilon \left\{ \left[ \begin{array}{l} \sum_{j \in \mathcal{F}} Z_j' F_j \\ \sum_{j \notin \mathcal{F}} \frac{1}{\theta(1-\theta)} (Y_j - R_j\theta) \end{array} \right] \Bigg| \; D,M,X \right\} ,$$

where $\varepsilon(\cdot \mid D,M,X)$ is the conditional expectation of the "full score function," given the observed data. Besides its obvious simplicity, this score function also has a simple interpretation; data from founders provide information about linkage-disequilibrium association, whereas nonfounders provide information about linkage. Recognizing that the preceding presentation of the general formulation is cumbersome, we have provided several key quantities in the score-estimating equation for a nuclear family, as an illustration:

The likelihood function of the observed nuclear-family data is as follows:

$$\begin{aligned} f(D,M,X) &\propto \sum_{g_1,g_2 p_1,p_2} f(d_1|g_1,x_1) f(d_2|g_2,x_2) f(g_1,p_1,m_1) f(g_2,p_2,m_2) \\ &\qquad \prod_{j=3}^{n} \sum_{g_j,p_j} f(d_j|g_j,x_j) f(g_j,p_j,m_j|g_{[j]},p_{[j]},m_{[j]}) \\ &= \sum_{g_1,g_2 p_1,p_2} f(d_1|g_1,x_1) f(d_2|g_2,x_2) f(m_1|g_1,p_1) f(m_2|g_2,p_2) f(g_1) f(g_2) \\ &\qquad \prod_{j=3}^{n} \left[ \sum_{g_j,p_j} f(d_j|g_j,x_j) f(r_{jp},y_{jp},r_{jm},y_{jm}) \right] , \end{aligned}$$

where, in the simulation studies,

$f(d_j|g_j,x_j)$ is the penetrance function for the $j$th individual, $j = 1,\ldots,6$ ;

$f(m_j|g_j,p_j) = [f(m_{j1}|g_{j1})f(m_{j2}|g_{j2})]^{p_j}[f(m_{j2}|g_{j1})f(m_{j1}|g_{j2})]^{\bar{p}_j}$, for founder $j = 1,2$ ;

$f(m_{jl}|g_{jl}) = [P(m_{jl} = 1|g_{jl})]^{m_{jl}}[P(m_{jk} = 0|g_{jl})]^{1-m_{jl}} = \dfrac{\exp[m_{jl}(\alpha + \beta g_{jl})]}{1 + \exp(\alpha + \beta g_{jl})}$; $l = 1,2$ ;

$f(g_j) = f(g_{j1})f(g_{j2})$, a product of two binomial probabilities for paired alleles ;

$f(r_{j1},y_{j1},r_{j2},y_{j2}) = [\theta_1^{y_1}(1-\theta_1)^{1-y_{j1}}]^{r_{j1}}[\theta_2^{y_{j2}}(1-\theta_2)^{1-y_{j2}}]^{r_{j2}}$ ;

$f(r_{j1},y_{j1},r_{j2},y_{j2}) = \theta^{y_{j1}r_{j1}+y_{j2}r_{j2}}(1-\theta)^{(1-y_{j1})r_{j1}+(1-y_{j2})r_{j2}}$, if $\theta_1 = \theta_2 = \theta$ .

The probability of the observed data, for putative paternally derived variables, is as follows (the probability for maternally derived variables is expressible in similar form):

$$f(g_1,p_1,D,M,X) = \sum_{g_2,p_2} f(d_1|g_1,x_1)f(d_2|g_2,x_2)f(m_1|g_1,p_1)f(m_2|g_2,p_2)f(g_1)f(g_2)\prod_{j=3}^{n}\left[\sum_{g_j,\ p_j} f(d_j|g_j,x_j)f(r_{jp},y_{jp},r_{jm},y_{jm})\right] .$$

After computing this joint distribution, one can immediately obtain the conditional distribution

$$f(g_1,p_1|D,M,X) = \frac{f(g_1,p_1,D,M,X)}{f(D,M,X)} ,$$

which is useful for several different computations.

The probability of the observed data, for the putative variable for the $j$th child, is as follows:

$$f(g_j,p_j,D,M,X) = \sum_{g_i,g_2,p_1,p_2} f(d_1 \mid g_1,x_1)f(d_2 \mid g_2,x_2)f(m_1 \mid g_1,p_1)f(m_2 \mid g_2,p_2)f(g_1)f(g_2)$$

$$[f(d_j|g_j,x_j)f(r_{jp},y_{jp},r_{jm},y_{jm})]\prod_{j'\neq j}^{n}\left[\sum_{g_{j'},p_{m_{j'}}} f(d_{j'}|g_{j'},x_{j'})f(r_{j'p},y_{j'p},r_{j'm},y_{j'm})\right] .$$

The paternal contribution to linkage disequilibrium in the estimating function is as follows:

$$u(\beta) = \sum_{g_1,p_1} u_\beta(d_1,g_1,x_1,p_1,m_1)f(g_1,p_1|D,M,X) .$$

where $u_\beta(d_1,g_1,x_1,p_{m_1},m_1) = Z_1'F_1$ represents the paternal contribution to $\beta$, if all putative factors have been observed. A similar expression is obtainable for maternal contribution, by replacement of $u_\beta(d_1,g_1,x_1,p_1,m_1)$ by $u_\beta(d_2,g_2,x_2,p_2,m_2)$.

The contribution to the recombination in the estimating function is as follows:

$$u(\theta) = \sum_{g_1g_2p_1p_2} f(d_1|g_1,x_1)f(d_2|g_2,x_2)f(m_1|g_1,p_1)f(m_2|g_2,p_2)f(g_1)f(g_2)$$

$$\left[\sum_{g_j,p_j} u_\theta(d_j,g_j,x_j,p_j,m_j)f(d_j|g_j,x_j)f(r_{jp},y_{jp},r_{jm},y_{jm})\right]\prod_{j'\neq j}^{n}\left[\sum_{g_{j'},p_{m_{j'}}} f(d_{j'}|g_{j'},x_{j'})f(r_{j'p},y_{j'p},r_{j'm},y_{j'm})\right] ,$$

where $u_\theta(d_j,g_j,x_j,p_j,m_j)$ represents the contribution to $\theta$ by the $j$th child, if all putative factors have been observed, and may be expressed as

$$u_\theta(d_j,g_j,x_j,p_j,m_j) = [\theta(1-\theta)]^{-1}[Y_j - R_j\theta] .$$

## References


Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, et al (1996) Accessing genetic information with high density DNA arrays. Science 274:610–614

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29: 311–322

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, et al (1987) A genetic linkage map of the human genome. Cell 51:319–337

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

Guo SW (1997) Linkage disequilibrium measures for fine-scale mapping: a comparison. Hum Hered 47:301–314

Haldane JBS (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet 8:299–309

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17:21–24

Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56: 519–527

Kruglyak L, Lander ES (1995) High-resolution genetic mapping of complex traits. Am J Hum Genet 56:1212–1223

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Lipshutz R (1997) Affymetrix GeneChip technology enables efficient access to complex genetic information. Genome Digest 4:10–11

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61: 439–448

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, Quillen J, et al (1994) A comprehensive human linkage map with centimorgan density: Cooperative Human Linkage Center (CHLC). Science 265: 2049–2054

Ott J (1989) Analysis of human genetic linkage. Johns Hopkins University Press, Baltimore

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402–409

Self SG, Longton G, Kopecky KJ, Liang KY (1991) On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53–61

Silverman BW (1990) Density estimation for statistics and data analysis, 1st ed. Chapman & Hall, New York

Spielman RS, McGinnis RE, Ewens WJ (1994) The transmission/disequilibrium test detects cosegregation and linkage. Am J Hum Genet 54:559–560

Terwilliger JD, Shannon WD, Lathrop GM, Nolan JP, Goldin LR, Chase GA, Weeks DE (1997) True and false positive peaks in genomewide scans: applications of length-biased sampling to linkage mapping. Am J Hum Genet 61:430–438

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Whittemore AS (1996) Genome scanning for linkage: an overview. Am J Hum Genet 59:704–716

Xiong M, Guo S-W (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531

Zhao LP, Grove J, Quiaoit F (1992) A method for assessing patterns of familial resemblance in complex human pedigrees, with an application to the nevus-count data in Utah kindreds. Am J Hum Genet 51:178–190

Zhao LP, Prentice RL (1990) Correlated binary regression using a quadratic exponential model. Biometrika 77:642–648

Zhao LP, Quiaoit F, Aragaki C, Hsu L. An efficient, robust and unified method for mapping complex traits (II): multipoint linkage analysis. Am J Med Genet (in press)

Zhao LP, Quiaoit F, Hsu L, Aragaki C (1998) An efficient, robust and unified method for mapping complex traits (I): two-point linkage analysis. Am J Med Genet 77:366–383