DOI: 10.1534/genetics.103.021600

Identifying the Susceptibility Gene(s) in a Set of Trait-Linked Genes Using Genotype Data

Ao Yuan,*,[1] Guanjie Chen,[†] Yuanxiu Chen,[†] Charles Rotimi[†] and George E. Bonney*

**Statistical Genetics and Bioinformatics Unit, [†]Genetic Epidemiology Unit, National Human Genome Center, Howard University, Washington, DC 20059*

Manuscript received August 25, 2003
Accepted for publication March 22, 2004

ABSTRACT

There are generally three steps to isolate a disease linkage-susceptibility gene: genome-wide scan, fine mapping, and, last, positional cloning. The last step is time consuming and involves intensive laboratory work. In some cases, fine mapping cannot proceed further on a set of markers because they are tightly linked. For years, genetic statisticians have been trying different ways to narrow the fine-mapping results to provide some guidance for the next step of laboratory work. Although these methods are practical and efficient, most of them are based on IBD data, which usually can be inferred only from the genotype data with some uncertainty. The corresponding methods thus have no greater power than one using genotype data directly. Also, IBD-based methods apply only to relative pair data. Here, using genotype data, we have developed a statistical hypothesis-testing method to pinpoint a SNP, or SNPs, suspected of responsibility for a disease trait linkage among a set of SNPs tightly linked in a region. Our method uses genotype data of affected individuals or case-control studies, which are widely available in the laboratory. The testing statistic can be constructed using any genotype-based disease-marker disequilibrium measure and is asymptotically distributed as a chi-square mixture. This method can be used for singleton data, relative pair data, or general pedigree data. We have applied the method to simulated data as well as a real data set; it gives satisfactory results.

RECENTLY, genome-wide scans have been widely used in the study of complex genetic diseases such as cardiovascular diseases, obesity, diabetes, schizophrenia, etc., due to the advance in biological science that hundreds of markers could be genotyped quickly with reduced cost. Subsequent fine-mapping studies have been also frequently reported, which narrow the linkage region to a disease trait to about one or a few centimorgans. However, very few of the studies reach the final step of positional cloning to isolate the gene responsible for the linkage to a complex disease. Part of the reason is that the process involves genomic DNA spanning millions of base pairs at the linkage region, sequencing large amounts of the overlapped genomic DNA fragments, and genotyping tens or hundreds of markers in the region, which take intensive work in an ordinary laboratory. In some cases, fine mapping cannot proceed farther on a set of markers because they are tightly linked. For years, genetic statisticians have been trying to develop parametric and/or nonparametric methods to pinpoint the linkage to one or very few markers suspected to be truly responsible for the linkage of a disease trait and to exclude those only in linkage disequilibrium (LD) to the susceptibility markers.

Difficulty in the identification of specific disease-predisposing alleles may result due to multiple genetic factors (Tait and Harrison 1991; Thomson 1991). Greenberg (1993) and Hodge (1993) considered the analysis of "necessary" *vs.* "susceptibility" loci in which the associated marker allele itself increases disease susceptibility but is neither necessary nor sufficient for disease expression. The conditioning method is one of the typical statistical tools for studying such problems. Fulker *et al.* (1999) developed a conditioning method using the variance component model. This method tests both linkage and association at the same time, so that it provides the result whether a locus is the candidate locus to the trait or is just in LD with the candidate locus. This idea was further expanded by Cardon and Abecasis (2000), in which a combined linkage and association method using the variance components model is proposed. Valdes and Thomson (1997) and Siegmund *et al.* (2001) used the conditioning method to narrow down the association region. Lazzeroni and Lange (1998) proposed such a framework in the transmission/disequilibrium test. Furthermore, Soria *et al.* (2000) considered a conditioning argument to pinpoint the linkage of the G20210A mutation in the prothrombin gene to the disease gene. On the other hand, Blanger *et al.* (2000)

[1]*Corresponding author:* Statistical Genetics and Bioinformatics Unit, National Human Genome Center, 2216 Sixth St., NW, Suite 205, Howard University, Washington, DC 20059.
E-mail: ayuan@howard.edu

studied a Bayesian variance components method, and Horikawa *et al.* (2000) used a modified association study method, which identified a single-nucleotide polymorphism (SNP), SNP43, that showed significant association with the evidence for linkage with type 2 diabetes.

Recently, Sun *et al.* (2002) proposed a statistical method for this problem. They used a conditioning hypothesis-testing procedure to pinpoint, among a set of tightly trait-linked genes, a single or a few susceptible markers, using identity-by-descent (IBD) data from affected sibships. This method is based on the genome-wide scan result, which identified a region showing strong linkage with a putative trait. Often markers in such a region are tightly linked among themselves. The goal of the method is to identify which of those markers are truly responsible for the linkage and which are merely tightly linked to such putative markers. This method is practical in application and yielded good results in their simulation studies.

However, most of the existing methods for this problem use IBD data on paired family members. Usually IBD data are not fully available in practice and can be inferred only from genotype data with uncertainty and often inconsistently from different methods used. Inference based on them has no greater power than that based on genotype data, unless the IBD data are a sufficient statistic for the parameters underlying the model. Also, IBD-based methods apply only to relative pair data.

Here we present a method for this problem by formulating a set of conditional hypothesis testing, in this respect similar to that in Sun *et al.* (2002), but we use genotype data instead and the testing statistic is different in nature from theirs. Using any genotype-based trait-marker disequilibrium measure, the testing statistics are constructed by successively conditioning on each of the tightly linked SNP sites. Our method is nonparametric: it does not require model specification or phase information in the data. It applies to family data of arbitrary structure, including singleton data, in which each individual comes from a different and independent family. Under the null hypothesis of being the sole susceptible site, each of these statistics follows asymptotically a chi-square mixture distribution. The corresponding *P* values are easily obtained via simulation.

THE METHOD

The data: Let A be the unknown disease allele, for which we want to infer its position in the human genome. Assume that there are J identified SNP markers, M_j ($j = 1, \ldots, J$), with alleles M_{jk} ($k = 1, 2$), which are brought to our attention because of their tight linkage to the disease allele. A natural question is whether all of them are susceptible genes for the linkage or some of them show disease linkage just because of their strong linkage with the true susceptible gene(s). Our goal here is to identify the true susceptible SNP(s), if any, among them.

For ease of explanation we first describe our method for singleton data and then extend it to general pedigree data in a later section. We now describe a general procedure for the conditional inference of this problem; the construction of the specific testing statistic is detailed later. Let $G = (G_1, \ldots, G_J)$ be a general notation for the composite SNP genotype at all the SNP loci, where $G_j = (g_{j1}, g_{j2})$ is its allelic notation; $G_{nj} = (g_{nj1}, g_{nj2})$ be the observed genotype of the nth individual at the jth SNP locus ($n = 1, \ldots, N$; $j = 1, \ldots, J$); and $G_n = (G_{n1}, \ldots, G_{nJ})$ be the vector notation of the observed composite genotype of individual n. The data to be used are $G_1, \ldots, G_N$, the observed composite genotypes of N individuals at J SNP loci each.

Here we assumed the common practice that at each SNP locus there are two different alleles in the population; we code them as 1 and 2, although the same value from alleles at different loci may have different allelic meaning. At each locus, we code the genotype as $G_{nj} = 0$ when $g_{nj1} \neq g_{nj2}$, $G_{nj} = I$ when $(g_{nj1}, g_{nj2}) = (1, 1)$, and $G_{nj} = II$ when $(g_{nj1}, g_{nj2}) = (2, 2)$. Note that we have two representations of a SNP genotype, one allelic and one numerical. Which one(s) will be used, even in the same expression, depends on convenience.

The disequilibrium measure and the conditioning principle: The proposed conditional testing procedure uses testing statistics, which are constructed via the conditional version of any trait-marker LD measure using genotype data. We first state the conditional testing principle and then give the specific forms of the testing statistic for some particular data designs.

Now we describe the trait-marker LD measure. Let p_A be the population frequency of the disease allele A, q_{jk} be those of allele k of marker j, and $P_{A,jk}$ be those of the haplotype (A, M_{jk}). Let $D_{A,jk} = P_{A,jk} - p_A q_{jk}$ be the LD measure between the disease allele A and allele k of marker j. Since the position of A is unknown, p_A, $P_{A,jk}$, and thus $D_{A,jk}$ cannot be directly estimated from the observed data; instead various quantities are constructed to infer it.

When $D_{A,jk}$ is positive, the marker allele M_{jk} is more likely to be associated with the disease-susceptible allele A than would be expected by chance. The disequilibrium measures $D_{A,jk}$ are among the main tools for finding the association between a marker locus (loci) and the disease locus. There are numerous ways to construct inference statistic from the $D_{A,jk}$'s, some using relative pair IBD data at markers, and some using marker genotype data (Bengtsson and Thomson 1981; Lehesjoki *et al.* 1993; Feder *et al.* 1996; Nielsen *et al.* 1998). Here we develop the conditional version of the genotype-based method.

Let $q_{jk|i}$ be the population frequency of allele k of the jth SNP conditional on the ith SNP genotype ($k = 1, 2$). Let $P_{A,jk|i}$ be the population frequency of the disease-

TABLE 1

Genotype frequencies at two loci

	Locus 2			
Locus 1	(1, 1)	(1, 2)	(2, 2)	Total
(1, 1)	39	72	45	156
(1, 2)	70	101	57	228
(2, 2)	23	70	25	118
Total	132	243	127	502

SNP haplotype (A, M_{jk}) at the jth marker locus conditional on the ith SNP genotype and $P_{jr|i}$ be that of the homozygote SNP genotype r at the jth SNP locus conditional on the ith SNP genotype ($r = I, II$). We choose the conditional LD measure at the jth locus, given the ith SNP genotype, as

$$D_{j|i} = P_{Aj1|i} - p_A q_{j1|i}. \quad (1)$$

Note that $P_{Aj2|i} - p_A q_{j2|i} = -(P_{Aj1|i} - p_A q_{j1|i})$, so only one of the marker alleles is needed to define this disequilibrium. Our motivation to use the conditional LD measure is that if marker i is the sole susceptible site of linkage to the disease allele, then the genotype data from this site constitute a sufficient statistic for this measure, or, in other words, it will explain all the disequilibria in the region. Thus, conditioning on the site of interest, the disequilibria parameters $D_{j|i}$ vanish from the conditional distribution of the data, for all $j \neq i$.

In the following we explain what the conditioning actually means in practice. Suppose we have genotype data for 502 individuals on two SNP loci, each locus has two alleles, and each allele takes one of the two forms that we coded as 1 and 2. The genotype at each locus is thus represented as (1, 1), (1, 2) = (2, 1) and (2, 2). The supposed observed genotype frequencies for the two loci are given in Table 1.

By conditioning on the genotype at the first locus being (1, 1), we mean the subgroup of 156 individuals whose genotype at the first locus is (1, 1). Within this subgroup, the genotype at the second locus is denoted as locus 2/(1, 1) and similarly for conditioning on the first locus genotype being (1, 2) and (2, 2). Thus conditioning on the first locus genotype (1, 1), (1, 2), and (2, 2) separately, the data are divided into three nonoverlap subdata sets, and we obtain the genotype frequency of the second locus as shown in Table 2. Likewise, conditioning on the second locus genotypes separately, we get the genotype frequency of the first locus as shown in Table 3.

In conditional testing, test statistics are constructed with the data in Tables 2 and 3. For example, to test the hypothesis that locus 1 is the only susceptible site, then conditioning on it we obtain three subtables. If the hypothesis is true, the LD vanish on each of the subtables, and the test statistics constructed from them should manifest nonsignificance.

The hypotheses and testing statistics: We are interested in testing the null hypothesis H_i: among the set of markers, SNP marker i is the sole cause of the linkage to the disease locus. Here we assume background effects on the linkage are negligible; see the DISCUSSION for more details on this. Under H_i, $D_{j|i} = 0$ for all $j \neq i$. For each fixed i, testing statistics $S_{j|i}$ are constructed, usually a function of the empirical version $\hat{D}_{j|i}$ $(j \neq i)$, such that they tend to be small under H_i and large otherwise. H_i can be decomposed as $H_i = H_{i1} \oplus H_{i2}$, where H_{ik} is the hypothesis: genotype k at site i is the sole susceptibility SNP reasonable for the trait LD in the region. Note that when H_{ik} is rejected, we can conclude only that the SNP does not contribute to the LD in the region, that the single or multiple causal polymorphism may not be among those that are typed, or that there is more than one source of such contribution. By testing the sequence of $\{H_{ik}\}$, we can find a confidence set, which may consist of a single SNP or several SNPs, or it may be empty. This set may be more accurately inferred by testing the hypothesis of multiple SNPs as in a later subsection. Our method can be used to detect a more detailed local relationship by testing the more detailed hypothesis $H_{j|i}$, LD at site j is completely caused by site i, or even the finer hypothesis $H_{j|ik}$, LD at site j is completely caused by genotype k of site i.

These last hypotheses are inferred using the statistics $S_{j|ik}$, which are the corresponding versions of the $S_{j|i}$'s for the $H_{j|ik}$'s. For recessive disease, the conditional statistic notation $S_{j|ik}$ means $S_{j|G_i=k}$. The $S_{j|ik}$'s are constructed of the form $S_{j|ik} = nX^2_{j|ik}$, and the random column vector $\sqrt{n}X_{ik}: = \sqrt{n}(X_{j|ik} : j \neq i)$ is jointly asymptotic normal under H_{ik}. Let Σ_{ik} be the asymptotic variance matrix of X_{ik} and $\lambda = (\lambda_1, \ldots, \lambda_{J-1})$ be its eigenvalues. Usually Σ_{ik}

TABLE 2

Genotype frequencies at locus 2 conditioning on locus 1

Locus 2/(1, 1)				Locus 2/(1, 2)				Locus 2/(2, 2)			
(1, 1)	(1, 2)	(2, 2)	Total	(1, 1)	(1, 2)	(2, 2)	Total	(1, 1)	(1, 2)	(2, 2)	Total
39	72	45	156	70	101	57	228	23	70	25	118

TABLE 3

Genotype frequencies at locus 1 conditioning on locus 2

Locus 1/(1, 1)				Locus 1/(1, 2)				Locus 1/(2, 2)			
(1, 1)	(1, 2)	(2, 2)	Total	(1, 1)	(1, 2)	(2, 2)	Total	(1, 1)	(1, 2)	(2, 2)	Total
39	70	23	132	72	101	70	243	45	57	25	127

and thus λ can be estimated by their empirical version. The particular forms of the $S_{j|ik}$'s are given later for different data designs.

Asymptotic distribution of the testing statistic: Let us consider H_{ik}. Its testing statistic is given by

$$S_{+|ik}(\lambda) = \sum_{j\neq i} \frac{S_{j|ik}}{\lambda_j} \quad \text{or} \quad S_{+|ik} = \sum_{j\neq i} S_{j|ik}.$$

To get the asymptotic distribution of $S_{+|ik}(\lambda)$ or $S_{+|ik}$ under H_{ik}, we first give a general result for the distribution of quadratic form of normal random variables. The proof is given in the APPENDIX.

Proposition: Let $X = (X_1, \ldots, X_d)'$ be a nondegenerate normal random vector: $X \sim N(\mathbf{0}, \Sigma)$ (*i.e.*, $|\Sigma| \neq 0$), with eigenvalues $\lambda = (\lambda_1, \ldots, \lambda_d)$; A is a d-dimensional positive definite symmetric matrix with eigenvalues $\gamma = (\gamma_1, \ldots, \gamma_d)$; the λ_i's and the γ_i's keep the same order in the diagonalization. We have

i. $X'AX \sim \chi^2_d(\gamma, \lambda) := \gamma_1\lambda_1 Y_1^2 + \ldots + \gamma_d\lambda_d Y_d^2$,

where the Y_j^2's are independent and identically distributed (IID) χ_1^2 random variables.

ii. Let $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_d)$ and $\Lambda = (\lambda_1, \ldots, \lambda_d)$; then

$$X'(A^{1/2})'\Gamma^{-1}\Lambda^{-1}A^{1/2}X \sim \chi^2_d.$$

Especially, when $A = I_d$, we have

$$X'\Lambda^{-1}X = \frac{X_1^2}{\lambda_1} + \ldots + \frac{X_d^2}{\lambda_d} \sim \chi^2_d.$$

Remark:

1. The case Σ or A being degenerate is not of much interest and can be avoided easily in the construction of the testing statistic.
2. It requires that γ and λ be of the same order; this can be done using the same orthogonal matrix (matrices) in the diagonalization of Σ and A. More conveniently, since it actually used only the $\gamma_j\lambda_j$'s, they are just the eigenvalues of $A\Sigma$ (or ΣA).
3. Using i or ii is a matter of choice. i is simpler in forming the χ^2 statistic but not in computing the quantiles or P values, while the order of the $\gamma_j\lambda_j$'s does not matter. ii involves computing $A^{1/2}$ in forming the χ^2 statistic, and the order of $\gamma_j\lambda_j$ and that of X_j must match. In practice, this is not trivial; however, it is simpler in computing the quantiles or P values using the existing χ^2 tables.
4. Given γ and λ, the density of $\chi^2_d(\gamma, \lambda)$ can be derived by the multiple convolution formula, and thus its αth quantile and/or the P value of the observed statistic can be obtained. But, more conveniently, for a given level α, the αth quantile and/or the P value of the observed statistic can be consistently estimated by their empirical versions.

To sample from $\chi^2_d(\gamma, \lambda)$, we sample $Y_1^2, \ldots, Y_{J-1}^2$ from χ_1^2 independently; then $\gamma_1\lambda_1 Y_1^2 + \ldots + \gamma_d\lambda_d Y_d^2$ is a sample from $\chi^2_d(\gamma, \lambda)$.

The χ^2 linear combination is the general form of the quadratic form of normals. When the X_j's are independent, $\lambda_j = \text{Var}(X_j)$; when the X_j's are IID and $A = I_d$, $\chi^2_d(\gamma, \lambda) = \lambda_1\chi^2_d$. There are some other similar results about the quadratic form of normals (GRAYBILL and MARSAGLIA 1957; GOOD 1969; KHATRI 1980, 1982; ANDERSON and STYAN 1982). Our result is independent and not of the same formulations and conditions as the others.

Let the eigenvalues (in their original order) of Σ_{ik} be $\lambda = (\lambda_1, \ldots, \lambda_{J-1})$; by ii of the *Proposition*, we have (see APPENDIX):

Corollary: Under H_{ik}, asymptotically

$$S_{+|ik}(\lambda) \sim \chi^2_{J-1},$$

and

$$S_{+|ik} \sim \chi^2_{J-1}(\lambda) := \lambda_1 Y_1^2 + \ldots + \lambda_{J-1} Y_{J-1}^2,$$

where the Y_j^2's are IID χ_1^2 random variables.

Thus for given $0 < \alpha < 1$, the asymptotic level α test of H_{ik} is given by the rejection rule: the P value of the observed $S_{+|ik}$ is smaller than α, or $S_{+|ik} > Q_{J-1}(\lambda, \alpha)$, the αth quantile of the $\chi^2_{J-1}(\lambda)$ distribution.

Note that our method requires only the genotype information and allele counts at each locus. It does not require phase information in diploids; thus it is practical in applications.

In the following we give the specific forms of the $S_{+|ik}$'s [$S_{+|ik}(\lambda)$'s] under some commonly used settings; those of the $S_{+|ik}(\lambda)$'s are the same and are omitted.

Multiple susceptible loci: Our method can be extended to the case of multiple susceptible loci without conceptual difficulty, but with more involved computa-

tions. Consider the hypothesis $H_{i_1k_1,\ldots,i_rk_r}$ $(1 \le r < J)$ that the composite genotypes $(k_1, \ldots, k_r)$ at loci $(i_1, \ldots, i_r)$ are the true susceptible ones. The corresponding testing statistics $S_{j|i_1k_1,\ldots,i_rk_r}$ are constructed similarly as before. The only difference is now the inference set, the conditional data set, consisting of those individuals whose alleles at loci $(i_1, \ldots, i_r)$ are $(k_1, \ldots, k_r)$, and

$$S_{+|_1k_1,\ldots,i_rk_r} = \sum_{j\notin\{i_1,\ldots,i_r\}} S_{j|i_1k_1,\ldots,i_rk_r},$$

which is asymptotically $\chi^2_{J-r}(\lambda)$, and $\lambda = (\lambda_1, \ldots, \lambda_{J-r})$ is the eigenvalue of the asymptotic variance matrix $\Sigma_{i_1k_1,\ldots,i_rk_r}$, which is estimated the same way as the single susceptible locus case, but uses the current inference data set.

For fixed r, there are $J!(J-r)!/r!$ of such tests across different choices of loci combinations, and 2^r of such tests for each choice of loci combination. So the total number of tests will be $2^rJ!(J-r)!/r!$.

Note that the above construction of the testing statistic is general; its inference behavior depends on the particular statistic used. The general form of the testing statistic is asymptotically a chi-square mixture, which is centralized under H_{ik} and noncentralized otherwise. The functional form of the parameters of interest entering the noncentrality parameter in the chi-square mixture will explain the behavior of the test in terms of asymptotic power. We give more detail on this for specific tests used in the following sections.

AFFECTED INDIVIDUAL DATA

Now we explain how to construct the $S_{+|ik}$'s in this type of data. In the case $J = 1$, assume the two SNP alleles are M and $\overline{M}$, and let A be the disease allele. Let p_A, q_M, and P_{AM} be the population frequency of the alleles A and M and haplotype AM, respectively, and let $D_{AM} = P_{AM} - p_Aq_M$ be the LD. For clarity we first assume the disease is *recessive* and $P(\text{Affected}|AA) = 1$. Under these assumptions, Feder *et al.* (1996) and more specifically Nielsen *et al.* (1998) discovered the relationship

$$F_M = \frac{P_{MM|\text{Affected}} + P_{\overline{M}\,\overline{M}|\text{Affected}} - q^2_{M|\text{Affected}} - q^2_{\overline{M}|\text{Affected}}}{1 - q^2_M - q^2_{\overline{M}}} = \psi(1-\psi)D^2_{AM}/(\phi^2 q_M q_{\overline{M}}),$$

where ψ is the probability that an individual will exhibit the disease due to causes other than this locus, and ϕ is the prevalence of the disease in the population. This equality enables us to detect the marker-disease association by testing Hardy-Weinberg disequilibrium at the marker locus without using IBD information. In fact the connection between the marker allele frequencies and the marker-disease LD is kept if we use only the numerator in the above equality, and this will simplify the computation. That is,

$$P_{MM|\text{Affected}} + P_{\overline{M}\,\overline{M}|\text{Affected}} - q^2_{M|\text{Affected}} - q^2_{\overline{M}|\text{Affected}} = 2\psi(1-\psi)D^2_{AM}/\phi^2. \quad (2)$$

We derive a conditional version of (2) to serve our purpose.

Since all individuals are affected in this study, we drop off the index "Affected" to simplify the notations. We want to test the hypothesis H_{ik}: SNP type k at locus i is the sole cause of the LD in the region. Let $P_{jr|ik}$ be the population frequency of genotype r ($r = I, II$) of locus j given one's genotype being k at locus i, $q_{jr|ik}$ be that of allele r ($r = 1, 2$) at locus j given one's genotype being k at locus i, ψ_j be the probability that an individual will exhibit the disease due to causes other than locus j, and $D_{j|ik}$ be the disequilibrium corresponding to the conditional LD measure. Now the same derivation of (2) leads to

$$T_{j|ik} := P_{jI|ik} + P_{jII|ik} - q^2_{j1|ik} - q^2_{j2|ik} = 2\psi_j(1-\psi_j)D^2_{j|ik}/\phi^2. \quad (3)$$

Under H_{ik}, all association of SNP j is completely explained by genotype k of locus i; thus $D_{j|ik} = 0$ and hence $T_{j|ik} = 0$ $(j \ne i)$.

We comment that our method works for a general disease model; in this case $T_{j|ik}$ is still a function of $D_{j|ik}$ but the expression is more involved (see Nielsen *et al.* 1998, pp. 1533–1534), and under H_{ik} we still have $T_{j|ik} = 0$ $(j \ne i)$; hence the test is still valid. In this case, the power and error rate computation will be more involved. The same comment applies to the case-control section also.

Now we construct testing statistics for H_{ik} $(i = 1, \ldots, J)$. The consistent estimates $\hat{P}_{jr|ik}$ of $P_{jr|ik}$ and $\hat{q}_{jr|ik}$ of $q_{jr|ik}$ are given by

$$\hat{P}_{jr|ik} = \frac{1}{N_{ik}}\sum_{n=1}^{N_{ik}} I_{n,jr|ik} \quad (r = I, II),$$

where $N_{ik} = \sum_{n=1}^{N} I(G_{ni} = k)$ is the total number of individuals with the ith SNP genotype being k, and we rearrange them as the first, second, ..., and the N_{ik}th individual. $I_{n,jr|ik}$, $(= 0, 1)$ is the indicator that the nth individual among this set has genotype type r on the jth locus given he (she) has genotype k at locus i, and

$$\hat{q}_{j1|ik} = \frac{1}{N_{ik}}\sum_{n=1}^{N_{ik}} \frac{J_{n,j1|ik}}{2}, \quad \hat{q}_{j2|ik} = 1 - \hat{q}_{j1|ik},$$

where $J_{n,j1|ik}(= 0, 1, 2)$ is, for the nth individual, the number of times allele 1 occurs at locus j, given one's genotype being k at locus i. The estimate of $T_{j|ik}$ is

$$\hat{T}_{j|ik} = \hat{P}_{jI|ik} + \hat{P}_{jII|ik} - \hat{q}^2_{j1|ik} - \hat{q}^2_{j2|ik}.$$

Let $\hat{T}_{ik} = (\hat{T}_{j|ik} : j \ne i)$ be the $(J-1)$ dimensional column vector. Under H_{ik}, $\sqrt{N_{ik}}\hat{T}_{ik}$ is asymptotically $N(\mathbf{0}, \Sigma_{ik})$ for some matrix Σ_{ik} to be identified later. Let $\lambda = (\lambda_1, \ldots, \lambda_{J-1})$ be all the eigenvalues of Σ_{ik}, and $S_{j|ik} = N_{ik}\hat{T}^2_{j|ik}$. By the *Corollary*, under H_{ik} asymptotically

$$S_{+|ik} = \sum_{j\ne i}\hat{S}_{j|ik} = N_{ik}\hat{T}'_{ik}\hat{T}_{ik} = N_{ik}\sum_{j\ne i}\hat{T}^2_{j|ik} \sim \chi^2_{J-1}(\lambda),$$

and Σ_{ik} is estimated by

$$\hat{\Sigma}_{ik} = \hat{D}\hat{\Omega}\hat{D}' \tag{4}$$

(APPENDIX), where

$$\hat{\Omega} = \frac{1}{N_{ik} - 1}\sum_{n=1}^{N_{ik}} Z_{n|ik}Z'_{n|ik}, \quad \hat{D} = \oplus_{j\neq i}(1,\ 2 - 4\hat{q}_{j1|ik})$$

and

$$Z_{n|ik} = \left((I_{n,jI|ik} + I_{n,jII|ik} - \hat{P}_{jI|ik} - \hat{P}_{jII|ik},\ \frac{J_{n,j1|ik}}{2} - \hat{q}_{j1|ik}) : j \neq i\right).$$

Here $\oplus$ means matrix direct summation, which results in a $(J - 1) \times 2(J - 1)$ dimensional matrix. From $\hat{\Sigma}_{ik}$, we obtain the estimated eigenvalues $\hat{\lambda}$.

Similarly, for H_i, let $N_i = N_{i1} + N_{i2}$, $\alpha_r = N_{ir}/N_i$,

$$T_{j|i} = T_{j|i1} + T_{j|i2} \quad \text{and} \quad \hat{T}_{j|i} = \hat{T}_{j|i1} + \hat{T}_{j|i2},$$

and $\hat{T}_i = (\hat{T}_{j|i} : j \neq i)$. Let Σ_i be the asymptotic matrix of $\hat{T}_i$ and $\lambda = (\lambda_1, \ldots, \lambda_{2(J-1)})$ be all the eigenvalues of Σ_i. Note that $\hat{T}_i = \hat{T}_{i1} + \hat{T}_{i2}$, and $\hat{T}_{i1}$ and $\hat{T}_{i2}$ are independent, so under H_i, $\sqrt{N_i}\hat{T}_i$ is asymptotically $N(\mathbf{0}, \Sigma_i)$, with $\Sigma_i = \alpha_1^{-1}\Sigma_{i1} + \alpha_2^{-1}\Sigma_{i2}$. Its estimate is obtained as $\hat{\Sigma}_i = \alpha_1^{-1}\hat{\Sigma}_{i1} + \alpha_2^{-1}\hat{\Sigma}_{i2}$, and $\hat{\Sigma}_{ir}$ is constructed as before.

Let

$$S_{j|i} = N\hat{T}^2_{j|i}, \quad S_{+|i} = \sum_{j\neq i}\hat{S}_{j|i}.$$

Under H_i, $S_{+|i} \sim \chi^2_{J-1}(\lambda)$.

We remark that in the above the asymptotic variance matrices Σ_{jk} are estimated the same way as for the IID data. In general the familial data are not IID, and the above variance matrices are dealt with differently. Usually, in the positive dependent case, the asymptotic variance matrix will be larger, in the sense of generalized variance—the determinant of the variance matrix—and consequently will tend to have larger eigenvalues than the IID case, such as the singleton data case. In the case of homogeneous familial structure, more accurate estimates can be obtained. We study the above methods for general pedigree data in the extension section later.

In some of the existing methods for this problem, *e.g.*, SUN *et al.* (2002), the conditional IBD sharing statistics are computed at each site given the genotype at that site. In this way the statistic can test whether each of the sites is the sole susceptible site, but will not be able to find the more detailed relationship between sites when the null hypothesis of only one susceptible site is rejected, while our test statistic can be used to reveal more detailed relationship. If H_{ik} is accepted, it is reasonable to say that the connection between site j and the disease locus is due to genotype k of site i.

By the asymptotic normality of $\sqrt{N_{ik}}\hat{T}_{ik}$ and (3), when H_{ik} is false, $S_{+|ik}$ will be asymptotically a noncentral $\chi^2_{J-1}(\lambda, \mu)$, with noncentrality parameter

$$\mu = \frac{4}{\phi^4}\sum_{j\neq i}\psi_j^2(1 - \psi_j)^2 D^4_{j|ik}.$$

It is clear that H_{ik} is true if and only if $D_{j|ik} = 0$ $(j \neq i)$. In terms of μ, the null hypothesis is rephrased as H_{ik}: $\mu = 0$. For a given level α $(= P(\text{reject } H_{ik}|H_{ik} \text{ is true}))$, the parameters λ, ψ_js, ϕ, and the $D_{j|ik}$'s, the asymptotic power of the test is

$$\beta = P(S_{+|ik} \geq Q_{J-1}(\lambda, \alpha)).$$

Here $Q_{J-1}(\lambda, \alpha)$ is the αth quantile of the noncentral $\chi^2_{J-1}(\lambda, \mu)$ distribution, which can be simulated by the sampling method after the *Remark* of the *Proposition*, but with $Y_1, \ldots, Y_{J-1}$ independent, and Y_j from $N(\mu_j, 1)$ with $\mu_j = 2\psi_j(1 - \psi_j)D^2_{j|ik}/\phi^2$ $(j \neq i)$.

For this particular test statistic, since the power is an increasing function of μ, H_{ik} will be more accurately rejected when the $\psi_j(1 - \psi)_j$'s and the conditional $D_{j|ik}$'s are large, and ϕ_j is small or the disease is relatively rare. Likewise, H_{ik} will be more correctly accepted when the $\psi_j(1 - \psi_j)$'s and the conditional $D_{j|ik}$'s are small (*i.e.*, mainly explained by allele k of locus i), and the disease is relatively common.

Likewise, the error rate, the probability of false acceptance, is

$$P(S_{+|ik} < Q_{J-1}(\lambda, \alpha)) = 1 - \beta.$$

CASE-CONTROL DATA

Let $q_{M|A}$ and $q_{M|U}$ denote marker M population frequencies for the affected (case) and unaffected (control) individuals. BENGTSSON and THOMSON (1981) and LEHESJOKI *et al.* (1993) gave the following LD measure:

$$R = \frac{q_{M|A} - q_{M|U}}{1 - q_{M|U}} = \frac{(1 - \psi)p_A D_{AM}}{\phi(1 - \phi)[q_{\bar{M}} + (1 - \psi)p_A D_{AM}/(1 - \phi)]}.$$

The conditional version of the above is

$$R(jr|ik) = \frac{q_{jr|A,ik} - q_{jr|U,ik}}{1 - q_{jr|U,ik}} = \frac{(1 - \psi)p_A D_{jr|ik}}{\phi(1 - \phi)[q_{\bar{jr}} + (1 - \psi)p_A D_{jr|ik}/(1 - \phi)]} \quad (r = 1, 2).$$

Let N_A and N_U be the number of affected and unaffected individuals, and

$$\hat{R}(jr|ik) = \frac{\hat{q}_{jr|A,ik} - \hat{q}_{jr|U,ik}}{1 - \hat{q}_{jr|U,ik}},$$

where

$$\hat{q}_{jr|A,ik} = \frac{1}{N_{A,ik}}\sum_{n=1}^{N_{A,ik}}\frac{J^A_{n,jr|ik}}{2}.$$

$N_{A,ik} = \Sigma_{n=1}^{N} I(A, G_{ni} = k)$ is the total number of "affected" individuals with the ith SNP genotype being k,

$$\hat{q}_{jr|U,ik} = \frac{1}{N_{U,ik}}\sum_{n=1}^{N_{U,ik}}\frac{J^U_{n,jr|ik}}{2},$$

where $J^A_{n,jr|ik}$ and $J^U_{n,jr|ik}$ are the same as the $J_{n,jr|ik}$ before, but here for affected and unaffected individuals. $N_{U,ik} =$

$\sum_{n=1}^{N} I(U, G_{ni} = k)$ is the total number of unaffected individuals whose ith SNP genotype is k. Let $N_{ik} = N_{A,ik} + N_{U,ik}$. Assume $N_{A,ik}/N_{ik} \to \alpha_{A,ik}$ and $N_{U,ik}/N_{ik} \to \alpha_{U,ik} = 1 - \alpha_{A,ik}$. To test H_{ik}, let

$$S_{j1|ik} = N_{ik}\hat{R}^2(j1|ik), \quad S_{+|ik} = \sum_{j\neq i} S_{j1|ik}.$$

Under H_{ik}, asymptotically

$$S_{+|ik} \sim \chi^2_{J-1}(\lambda),$$

where λ is the vector of eigenvalues of the matrix Σ_{ik}. Let

$$Z^{A}_{n|ik} = \left(\frac{J^{A}_{n,j1|ik}}{2} - \hat{q}_{j1|A,ik} : j \neq i\right),$$

$$Z^{U}_{n|ik} = \left(\frac{J^{U}_{n,j1|ik}}{2} - \hat{q}_{j1|U,ik} : j \neq i\right) \quad (k = 1, 2).$$

Then Σ_{ik} is estimated by

$$\hat{\Sigma}_{ik} = \hat{D}\hat{\Omega}\hat{D}' \tag{5}$$

(APPENDIX), where, for singleton data, the affected and the unaffected are independent, so $\hat{\Omega} = \alpha^{-1}_{A,ik}\hat{\Omega}_A \oplus \alpha^{-1}_{U,ik}\hat{\Omega}_U$,

$$\hat{\Omega}_A = \frac{1}{N_{A,ik} - 1}\sum_{n=1}^{N_{A,ik}} Z^{A}_{n|ik}(Z^{A}_{n|ik})',$$

$$\hat{\Omega}_U = \frac{1}{N_{U,ik} - 1}\sum_{n=1}^{N_{U,ik}} Z^{U}_{n|ik}(Z^{U}_{n|ik})',$$

and

$$\hat{D} = \oplus_{j\neq i}\left(\frac{1}{1 - \hat{q}_{j1|A,ik}}, \frac{\hat{q}_{j1|A,ik} - 1}{(1 - \hat{q}_{j1|U,ik})^2}\right).$$

Similarly, to test H_i, let

$$S_{+|i} = \sum_{k=1}^{2} S_{+|ik}.$$

Then under H_i, asymptotically $S_{+|i} \sim \chi^2_{J-1}(\lambda)$, and λ is the vector of eigenvalues of Σ_i, which is estimated by

$$\hat{\Sigma}_i = \hat{D}\hat{\Omega}\hat{D}' \tag{6}$$

(APPENDIX), where, for singleton data, $\hat{\Omega} = \alpha^{-1}_{A,i}\hat{\Omega}_A \oplus \alpha^{-1}_{U,i}\hat{\Omega}_U$,

$$\hat{\Omega}_A = \frac{1}{N_{A,i} - 1}\sum_{n=1}^{N_{A,i}} Z^{A}_{n|i}(Z^{A}_{n|i})' \text{ and } \hat{\Omega}_U = \frac{1}{N_{U,i} - 1}\sum_{n=1}^{N_{U,i}} Z^{U}_{n|i}(Z^{U}_{n|i})',$$

where $Z^{A}_{n|i} = (Z^{A}_{n|i1}, Z^{A}_{n|i2})$, $Z^{U}_{n|i} = (Z^{U}_{n|i1}, Z^{U}_{n|i2})$, $N_i = N_{i1} + N_{i2}$, $N_{A,i} = N_{A,i1} + N_{A,i2}$, $N_{U,i} = N_{U,i1} + N_{U,i2}$, $\alpha_{A,i} = N_{A,i}/N_i$, and $\alpha_{U,i} = N_{U,i}/N_i = 1 - \alpha_{A,i}$. Similarly,

$$\hat{D} = \oplus_{j\neq i,k=1,2}\left(\frac{1}{1 - \hat{q}_{j1|A,ik}}, \frac{\hat{q}_{j1|A,ik} - 1}{(1 - \hat{q}_{j1|U,ik})^2}\right).$$

Other LD measures can also be used, for example, the trend test statistic (ARMITAGE 1955; DEVLIN and ROEDER 1999).

As in the affected individual case, when H_{ik} is not true, $S_{+|ik}$ is asymptotically noncentral $\chi^2_{J-1}(\lambda, \mu)$, where

$$\mu = \frac{p_A^2}{\phi^2(1 - \phi)^2}\sum_{j\neq i}\frac{(1 - \psi_j)^2 D^2_{j1|ik}}{[q_{j2} + (1 - \psi_j)p_A D_{j1|ik}/(1 - \phi_j)]^2}.$$

Given α, λ, ψ_j, ϕ, p_A, the q_{j2}'s, and the $D_{j1|ik}$'s, the power and error rate can be computed by simulation as before, but with $Y_1, \ldots, Y_{J-1}$ independent, with Y_j from $N(\mu_j, 1)$, where

$$\mu_j = \frac{p_A(1 - \psi_j)D_{j1|ik}}{\phi(1 - \phi)[q_{j2} + (1 - \psi_j)p_A D_{j1|ik}/(1 - \phi)]} \quad (j \neq i).$$

Here, the power and probability of correct acceptance of H_{ik} depend on ψ, ϕ, p_A, the q_{j2}, and the $D_{j1|ik}$'s. The power is maximum when the conditional $D_{j1|ik}$'s are maximum, and the test is more likely to accept H_{ik} when the $D_{j1|ik}$'s are small. Their relationships with the other parameters can be analyzed similarly.

EXTENSION TO GENERAL PEDIGREE DATA

As mentioned earlier, the only difference in our methods between general pedigree data and the singleton data is the estimations of the corresponding asymptotic variance matrices. A simple method for this purpose can be found in the work of G. E. BONNEY, V. APPREY and A. YUAN (unpublished data), without any assumption on the data and no extra parameters introduced for the dependence. We illustrate this with the affected familial data, which for the case-control family data is similar. For such data, the estimations for the genotype/allele frequencies in the previous sections are not IID averages; we rewrite them as IID versions, so that their asymptotic variance matrices can be computed easily. First we assume the data have the same familial structure. Suppose there are M families with S individuals each ($N = MS$). We redefine $\hat{P}_{jr|ik}$ as

$$\hat{P}_{jr|ik} = \frac{1}{M_{ik}}\sum_{m=1}^{M_{ik}}\sum_{s=1}^{S}\frac{s}{S}I_{jr|ik}(s, m) \quad (r = I, II),$$

where $M_{ik} = \sum_{m=1}^{M}\sum_{s=1}^{S} I_{ik}(s, m)$ is the total number of families in which at least one individual with SNP type k at locus i, $I_{ik}(s, m)$, is the indicator that in the mth family, there are s individuals with SNP type k at locus i. $I_{jr|ik}(s, m)$ is the indicator that there are s individuals in family m with SNP type r on the jth locus, given the family is in the group with SNP type k on the ith locus. Let $I_{jr|ik}(m) = \sum_{s=1}^{S}(s/S)I_{jr|ik}(s, m)$. Then for fixed (jr, ik), $\{I_{jr|ik}(m) : m = 1, \ldots, M\}$ is an IID sequence, and for different (jr, ik) and $(j'r', i'k')$, $\{I_{jr|ik}(m) : m = 1, \ldots, M\}$ and $\{I_{j'r'|i'k'}(m) : m = 1, \ldots, M\}$ are independent. Similarly, $\hat{q}_{jr|ik}$ is redefined as

$$\hat{q}_{jr|ik} = \frac{1}{M_{ik}}\sum_{m=1}^{M_{ik}}\sum_{s=1}^{2S}\frac{s}{2S}J_{jr|ik}(s, m) \quad (r = 1, 2),$$

where $J_{jr|ik}(s, m)$ is the count that there are s SNP allele

r in family m on the jth locus, and their SNP type is k on the ith locus. Let $J_{jr|ik}(m) = \sum_{s=1}^{2S} (s/2S) J_{jr|ik}(s, m)$. Then for fixed (jr, ik), $\{J_{jr|ik}(m) : m = 1, \ldots, M\}$ is an IID sequence, and for different (jr, ik) and $(j'r', i'k')$, $\{J_{jr|ik}(m) : m = 1, \ldots, M\}$ and $\{J_{j'r'|i'k'}(m) : m = 1, \ldots, M\}$ are independent.

Let $\hat{T}_{j|ik}$ and $\hat{T}_{ik}$ be as before but with $\hat{P}_{jI|ik}$, $\hat{P}_{jII|ik}$, and $\hat{q}_{j1|ik}$ replaced by the above versions. Let $S_{+|ik} = M_{ik}\hat{T}'_{ik}\hat{T}_{ik}$. Now it is clear that the $\hat{\Omega}$ in (4) can be replaced by the consistent estimator for this case as

$$\hat{\Omega} = \frac{1}{M_{ik} - 1} \sum_{m=1}^{M_{ik}} Z_{m|ik} Z'_{m|ik},$$

where

$$Z_{m|ik} = \left(\left(I_{jI|ik}(m) + I_{jII|ik}(m) - \hat{P}_{jI|ik} - \hat{P}_{jII|ik}, \frac{J_{j1|ik}(m)}{2} - \hat{q}_{j1|ik} \right) : j \neq i \right)$$

and $\hat{\Sigma}_{ik} = \hat{D}\hat{\Omega}\hat{D}'$, and $\hat{D}$ is the same as in (4).

More generally, suppose that there are L different familial structures in the data set, with size M_l each, and the lth structure has S_l individuals per family ($l = 1, \ldots, L$). Let

$$\hat{P}^{(l)}_{jr|ik} = \frac{1}{M^{(l)}_{ik}} \sum_{m=1}^{M^{(l)}_{ik}} \sum_{s=1}^{S_l} \frac{s}{S_l} I^{(l)}_{jr|ik}(s, m) \quad (r = 1, 2;\ l = 1, \ldots, L),$$

where $M^{(l)}_{ik} = \sum_{m=1}^{M_l} \sum_{s=1}^{S_l} I^{(l)}_{ik}(s, m)$ is the total number of families with the structure l in which at least one individual with SNP type k at locus i, $I^{(l)}_{ik}(s, m)$ and $I^{(l)}_{jr|ik}(s, m)$, is the counterpart of $I_{ik}(s, m)$ and $jr|ik(s, m)$, respectively, for familial structure l. Let $I^{(l)}_{jr|ik}(m) = \sum_{s=1}^{S_l} (s/S_l) I^{(l)}_{jr|ik}(s, m)$. Then for fixed (l, jr, ik), $\{I^{(l)}_{jr|ik}(m) : m = 1, \ldots, M\}$ is an IID sequence, and for different (l, jr, ik) and $(l', j'r', i'k')$, $\{I^{(l)}_{jr|ik}(m) : m = 1, \ldots, M\}$ and $\{I^{(l')}_{j'r'|i'k'}(m) : m = 1, \ldots, M\}$ are independent. Let $M_{ik} = \sum_{l=1}^{L} M^{(l)}_{ik}$ define the estimate of $P_{jr|ik}$ as

$$\hat{P}_{jr|ik} = \sum_{l=1}^{L} \frac{M^{(l)}_{ik}}{M_{ik}} \hat{P}^{(l)}_{jr|ik} \quad (r = I, II).$$

Similarly, let

$$\hat{q}^{(l)}_{jr|ik} = \frac{1}{M^{(l)}_{ik}} \sum_{m=1}^{M^{(l)}_{ik}} \sum_{s=1}^{2S_l} \frac{s}{2S_l} J^{(l)}_{jr|ik}(s, m) \quad (r = 1, 2),$$

where $J^{(l)}_{jr|ik}(s, m)$ is the counterpart of $J_{jr|ik}(s, m)$ for familial structure l. Let $J^{(l)}_{jr|ik}(m) = \sum_{s=1}^{2S_l} (s/2S_l) J^{(l)}_{jr|ik}(s, m)$, and define

$$\hat{q}_{j1|ik} = \sum_{l=1}^{L} \frac{M^{(l)}_{ik}}{M_{ik}} \hat{q}^{(l)}_{j1|ik}.$$

Now for this general pedigree data, let $\hat{T}_{j|ik}$ and $\hat{T}_{ik}$ be as before but with $\hat{P}_{j1|ik}$, $\hat{P}_{j2|ik}$, and $\hat{q}_{j1|ik}$ replaced by the above versions. Let $S_{+|ik} = M_{ik}\hat{T}'_{ik}\hat{T}_{ik}$, and we assume $\alpha^{(l)}_{ik} := \lim M^{(l)}_{ik}/M_{ik} > 0$ ($l = 1, \ldots, L$), then a consistent estimate of Σ_{ik} is given by

$$\hat{\Sigma}_{ik} = \hat{D}\left(\sum_{l=1}^{L} \frac{M^{(l)}_{ik}}{M_{ik}} \hat{\Omega}_l \right) \hat{D}' \tag{7}$$

(APPENDIX), where

$$\hat{\Omega}_l = \frac{1}{M^{(l)}_{ik} - 1} \sum_{n=1}^{M^{(l)}_{ik}} Z^{(l)}_{n|ik} (Z^{(l)}_{n|ik})', \ \hat{D} = \oplus_{j \neq i} (1, 2 - 4\hat{q}_{j1|ik})$$

and

$$Z^{(l)}_{n|ik} = \left((I^{(l)}_{n,jI|ik} + I^{(l)}_{n,jII|ik} - \hat{P}^{(l)}_{jI|ik} - \hat{P}^{(l)}_{jII|ik}, \frac{J^{(l)}_{n,j1|ik}}{2} - \hat{q}^{(l)}_{j1|ik}) : j \neq i \right),$$

$$l = 1, \ldots, L.$$

For the test of H_i, or the case of case-control data, testing statistics and the corresponding asymptotic variance matrices can be obtained in a similar way; we omit the details here.

SIMULATION STUDY

Here we use simulated data to illustrate our method. To exhibit the applicability of our method, we use singleton data, which is out of the scope of the IBD-based methods. We simulate the data $G_1, \ldots, G_N$, where $G_n = (G_{n1}, \ldots, G_{nJ})$ ($n = 1, \ldots, N$) and $G_{nj} = (g_{nj1}, g_{nj2})$, the two alleles at SNP site j for the nth individual. The g_{njk}'s are coded as 1, 2 for its possible two alleles. We assume phase is known to simplify the simulation process, so that for each n, the two haplotypes $(g_{n11}, \ldots, g_{nJ1})$ and $(g_{n12}, \ldots, g_{nJ2})$ are independent. In this example, we take $J = 6$, so all the vectors $G_n = (G_{n1}, \ldots, G_{n6})$ are random samples from the population genotype $S = (S_1, \ldots, S_6)$, and $S_j = (s_{j1}, s_{j2})$ is the genotype at the jth site. We assume genotype (1, 1) at the third SNP site is responsible for all the LD with the disease allele A; the other first alleles, $s_{j1} (j \neq 3)$, in this region are tightly linked to s_{31}.

Now the haplotypes $S^{(1)} = (s_{11}, \ldots, s_{61})$ and $S^{(2)} = (s_{12}, \ldots, s_{62})$ are independent and the s_{j2}'s are independent within themselves. Denote $G^{(1)}_n = (g_{n11}, \ldots, g_{n61})$ and $G^{(2)}_n = (g_{n12}, \ldots, g_{n62})$ as the two haplotypes of the nth individual. To sample such data, for each n we need only to sample $G^{(1)}_n$ from $S^{(1)}$ and $G^{(2)}_n$ from $S^{(2)}$ independently. Let $q_A = 0.8$ be the frequency of the disease allele allele $A = 1$ among the affected individuals, $q^{(1)} = (q_{11}, \ldots, q_{61})$ be the frequencies of $S^{(1)} = (1, \ldots, 1)$, and $q^{(2)} = (q_{12}, \ldots, q_{62})$ be that of $S^{(2)} = (1, \ldots, 1)$. To sample from $S^{(2)}$ is trivial; *i.e.*, just sample g_{nj2} independently from $B(q_{j2})$, the Bernoulli distribution with probability q_{j2} of getting 1 and probability $1 - q_{j2}$ of getting 0. To sample $G^{(1)}_n$, we need to sample from a joint Bernoulli distribution with probability $q^{(1)}$. Such a joint distribution can be specified in the form

$$P(S^{(1)}) = \exp\{\Psi' S^{(1)} + \Omega' W - A(\Psi, \Omega)\}$$

(Cox 1972; Fitzmaurice and Laird 1993), where Ψ and Ω are parameters and $\exp\{-A(\Psi, \Omega)\}$ is the nor-

TABLE 4

Affected individual data: values of observed $S_{+|j1}$ (P value) for different q

q =	j = 1	j = 2	j = 3	j = 4	j = 5	j = 6
0.1	16.120 (0.000)	16.586 (0.000)	0.205 (0.639)	17.233 (0.000)	20.783 (0.000)	19.001 (0.000)
0.3	36.984 (0.000)	46.335 (0.000)	0.851 (0.440)	50.917 (0.000)	44.753 (0.000)	49.209 (0.000)
0.5	35.772 (0.000)	30.164 (0.000)	1.339 (0.311)	25.618 (0.000)	32.761 (0.000)	33.267 (0.000)
0.7	9.051 (0.000)	8.849 (0.000)	0.753 (0.562)	12.694 (0.000)	7.736 (0.000)	9.531 (0.000)
0.9	1.455 (0.0114)	1.184 (0.0218)	0.218 (0.420)	1.938 (0.0026)	0.299 (0.310)	0.264 (0.352)

malizing constant and W is all the cross-product terms of $S^{(1)}$, including all the second- and higher-order terms. This distribution can be sampled using the Gibbs sampler (Geman and Geman 1984). But the specification of the joint Bernoulli distribution has some subjectivity and the sampling scheme is not simple. Instead, we use a normal discretization method to sample it. We use high correlation for linkage. Let Σ be the corresponding correlation matrix of the $J+1$-dimensional normal distribution for $(A, S^{(1)})$,

$$\Sigma = \begin{pmatrix} 1 & & & 0.5 & & & \\ & 1 & & 0.5 & & & \\ & & 1 & 0.5 & & & \\ 0.5 & 0.5 & 0.5 & 1 & 0.34 & 0.3 & 0.21 \\ & & & 0.34 & 1 & & \\ & & & 0.3 & & 1 & \\ & & & 0.21 & & & 1 \end{pmatrix}.$$

Note that this matrix corresponds to a strong connection between A and s_{31}, but not between A and $(s_{11}, s_{21}, s_{31}, s_{41}, s_{51}, s_{61})$; it also corresponds to a strong connection between s_{31} and $(s_{11}, s_{21}, s_{31}, s_{41}, s_{51}, s_{61})$. Thus all the loci have apparent linkage with the disease allele A.

To sample the composite genotypes from the above distribution, let $X = (x_A, x_1, \ldots, x_6)$ be a sample from the normal distribution $N(\mathbf{0}, \Sigma)$; if $x_j < \Phi^{-1}(q_{j1})$, we assign $g_{nj1} = 1$; otherwise $g_{nj1} = 0$, $(j = 1, \ldots, 6)$, where $\Phi^{-1}(q)$ is the qth quantile of the standard normal distribution. Since q_{31} is the proportion of allele 1, at locus 3, which is linked to the disease allele, in the affected population, the two alleles at locus 3 are in Hardy-Weinberg disequilibrium. The disease is recessive. We make the corresponding conditional probability $P(s_{32} = 1 | s_{31} = 1)$ high, say 0.8, among the affected individuals. In the simulation, we used a high frequency of $q_{j1} = q = 0.1, 0.2, \ldots, 0.9$, $(j \neq 3)$ for allele 1 at each locus, to see how this affects the results.

By the same way we simulated control data, in which the two haplotyes are sampled the same way as $G_n^{(2)}$ above. Together with the previous affected data we have case-control data, and the analysis is displayed in Table 6.

Specifically, the sampling scheme has the following three steps:

For each $n = 1, \ldots, N$, $(N = 1000)$:

i. Draw a sample $X = (x_A, x_1, \ldots, x_J)$ from the normal distribution $N(\mathbf{0}, \Sigma)$; if $x_j < \Phi^{-1}(q_{j1})$, we assign $g_{nj1} = 1$; otherwise $g_{nj1} = 0_3$ $(j = 1, \ldots, 6)$. Then we get the sample $G^{(1)} = (g_{n11}, \ldots, g_{nJ1})$.
ii. If $g_{n31} = 1$, set $q_{32} = P(s_{32} = 1 | s_{31} = 1) = 0.8$, else $q_{32} = 0.1$. For each $j = 1, \ldots, J$, draw X from $U(0, 1)$, the uniform distribution on $[0, 1]$; if $X < q_{j2}$ assign $g_{nj2} = 1$; otherwise assign $g_{nj2} = 0$. Then we get a sample $G^{(2)} = (g_{n12}, \ldots, g_{nJ2})$.
iii. $G_n = (G^{(1)}, G^{(2)})$ is a sample from S.

When the two alleles at each locus are in Hardy-Weinberg disequilibrium, we use a two-dimensional normal with mean (0, 0) and variance matrix $\Omega = (1, r; r, 1)$ with $r = 0.2$ to model their dependence. For each n, we first get the sample $G^{(1)} = (g_{n11}, \ldots, g_{nJ1})$ from $(x_1, \ldots, x_J)$ as before, then for each $j = 1, \ldots, J$ separately, sample y_j from the conditional distribution $N(rx_j, 1 - r^2)$. If $y_j < \Phi^{-1}(q_{j2})$, assign $g_{nj2} = 1$, otherwise 0.

To simulate the case-control data, we choose $q = 0.6$ for the case and $q = 0.25$ for the control.

RESULTS

Simulated data: We constructed the test statistics $S_{+|ik}$, $(i = 1, \ldots, J; k = 1, 2)$ and computed the corresponding eigenvalues $\lambda = (\lambda_1, \ldots, \lambda_{J-1})$, using the method described in the *Remark* after the *Proposition* to compute the χ^2 P value under the null hypotheses. Since in the simulation the sole linkage with the disease allele comes from s_{31}, we expect H_{31} will be accepted, and the other H_{jk}'s will be rejected. Table 4 is a summary of the observed values of the $S_{+|j1}$'s for the H_{jk}'s, for different choices of q, with corresponding P values in parentheses. We simulated and computed data for $q = 0.1, 0.2, \ldots, 0.9$; we display only part of them to save space.

For each testing statistic $S_{+|jk}$, there is a set of nonnegative eigenvalues $\lambda = (\lambda_1, \ldots, \lambda_{J-1})$. Their magnitude plays an important role in determining the asymptotic P value of the observed $S_{+|jk}$. For a given observed value of $S_{+|jk}$ and fixed number of loci J, a roughly larger eigenvalue total $|\lambda|$ (defined as $\lambda_1 + \ldots + \lambda_{J-1}$) results in a larger P value, and vice versa. Although for two sets of eigenvalues $\lambda_1 = (\lambda_{11}, \ldots, \lambda_{1,J-1})$ and $\lambda_2 = (\lambda_{21}, \ldots, \lambda_{2,J-1})$, even if $|\lambda_1| = |\lambda_2|$, the corresponding distributions

TABLE 5

Affected individual data: values of observed $S_{+|j1}$ (P value) for different r

$r =$	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	$j = 6$
0.05	10.922 (0.000)	12.735 (0.000)	1.371 (0.246)	12.479 (0.000)	11.419 (0.000)	14.138 (0.000)
0.1	8.524 (0.000)	11.361 (0.000)	1.390 (0.200)	13.723 (0.000)	8.770 (0.000)	9.066 (0.000)
0.2	15.131 (0.000)	14.726 (0.000)	3.186 (0.007)	17.919 (0.000)	15.174 (0.000)	10.718 (0.000)

$\chi^2(\lambda_1)$ and $\chi^2(\lambda_2)$ may not be equal, and they are equal if and only if $\lambda^{(1)} = \lambda^{(2)}$, where $\lambda^{(k)} = (\lambda_{k,(1)}, \ldots, \lambda_{k,(J-1)})$ is the ordered version of λ_k ($k = 1, 2$).

We display in the following the eigenvalues $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{j5})$ for the $S_{+|j1}$'s, for the case $q = 0.7$.

$$\lambda_1 = (0.25, 0.23, 0.19, 0.17, 0.14), \quad \lambda_2 = (0.24, 0.22, 0.17, 0.15, 0.13),$$

$$\lambda_3 = (0.23, 0.21, 0.19, 0.18, 0.17), \quad \lambda_4 = (0.25, 0.21, 0.18, 0.16, 0.11),$$

and

$$\lambda_5 = (0.24, 0.21, 0.19, 0.15, 0.13), \quad \lambda_6 = (0.23, 0.20, 0.16, 0.15, 0.13).$$

We find that in most cases the P values of $S_{+|31}$ suggest acceptance of H_{31} with high confidence, and those for $S_{+|j1}$ ($j \neq 3$), suggest rejection of H_{j1}, except for the case $q = 0.9$, in which the P values of $S_{+|51}$ and $S_{+|61}$ are also significant, along with that of $S_{+|31}$. We regard this last case as exceptional, in which the over-high proportion of allele 1 at each locus blurred the identifiability of the problem (think of the extreme case of $q \approx 1$; the corresponding locus contributes nearly no information for the problem). Thus, in all these cases, the true hypothesis H_{31} is accepted with high confidence, and the other false ones, H_{j1}, are rejected; *i.e.*, the true disease-linkage-related allele 1 at locus 3 is correctly identified among all six loci that are all in LD with the disease locus.

To investigate the influence of the deviation from Hardy-Weinberg on our method, we simulated the data for this case, in which we use the allelic correlation $r \neq 0$ at each locus for the deviation from Hardy-Weinberg equilibrium (HWE). The disease allele population frequency is fixed at $q = 0.7$ and the results are displayed in Table 5.

In the non-HWE case, it seems that the true picture becomes more difficult to recover as the deviation from HWE increases. In general, significant departures from HWE are not expected, but if observed, caution should be taken in applying this method (if genotyping error is present, for example). In particular, in situations in which nonrandom mating is a known confounder because of inbreeding or population structure, care should be exercised.

For the case-control data, we used $q = 0.6$ for the case and $q = 0.25$ for the control; HWE is assumed, and again locus 3 is the only connection to the disease allele. The results are shown in Table 6. It is seen that again, for the case-control data SNP locus 3 is correctly identified, and all the other loci are rejected as sources of cause for LD in the region.

The following is a tabulation of power of the test for the above simulated data, using the above λ and some combinations of α, $\psi_j = \psi (j \neq i)$, ϕ, and $D_{j|31} = d (j \neq 3)$. To get a sense of the power behavior of our methods, we choose $J = 6$, $\lambda = \lambda_1$ as shown before. The noncentrality parameter μ involves $2J - 1$ parameters in it. It is impractical to investigate and tabulate the influence of each of the $2J - 1$ parameters to the power. Instead, we investigate the influence of μ to the power, with the given genetic model. Each given value of μ, corresponding to a $2(J - 1)$-dimensional parameter subspace, is given by the formula for μ. Table 7 shows the display of power for both the affected individual data and the case-control data, for some choices of the level α and the parameter μ. We comment that for the above specification of the parameter μ, the power of the tests for both the affected individual data and that for the case-control data are the same.

Since the μ in the power of the test for affected individual data and that for the case-control data have different expressions, more detailed power computation can be obtained by the specification in terms of all the parameters involved.

Application to real data: *Non-insulin-dependent diabetes mellitus-1 data:* We first apply our method to the non-insulin-dependent diabetes mellitus-1 (NIDDM1) data

TABLE 6

Case-control data: values of observed $S_{+|j1}$ (P value)

$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	$j = 6$
2.2 (0.0016)	1.7 (0.0001)	0.16 (0.160)	3.3 (0.0001)	4.2 (0.000)	5.1 (0.000)

TABLE 7

Power for some given parameter values

α/μ	0.1	0.5	1.0	1.5	2.0
0.01	0.0182	0.1020	0.4536	0.8250	0.9820
0.02	0.0332	0.1648	0.5684	0.8734	0.9880
0.05	0.0934	0.2868	0.6987	0.9432	0.9977

used in SUN *et al.* (2002) and list our results along with theirs in Table 8. We see that, for these data, the two methods yield quite different, although not contradictory, results. With the method of Sun *et al.*, loci 2 and 12 are most likely responsible for the LD, while by our method, loci 2, 4, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, 19, 20, and 22 all likely contribute to the LD in the region. One possibility for the difference of the two methods might be that the calpain-10 region has some patterns of LD that are not understood—violating one of the assumptions of the methods. Since the truth in the data is unknown, we do not comment on the performances of the two methods on these data. It is not uncommon in the hypothesis test context, even for methods based on the same type of data, that different methods may have different results, even contradictory ones. In principle, methods using genotype data have no less power in inference than those using IBD data. Here it is too early to comment on the pros and cons for the two types of methods. A formal assessment may involve long-term and large-scale studies. At least our method provides the user more options and a flexible tool for this problem. Also, more methods will give us more strength in the inference. If the methods give consistent results, this will strengthen our confidence in decision; if they do not or are contradictory, the problem may need further investigation. We may perform the hypothesis tests on the current confidence set and continue this way to get a final confidence set of SNPs, in which all of them are accepted as possible sources of LD in the region. We do not pursue this in detail here because of space limitation.

Diabetes data: Next, in a diabetes study, 280 individuals with type 2 diabetes were genotyped at a large number of SNP sites. First we find those SNPs with strong linkage to the trait and then use our method to identify the susceptible one(s). We use the measure of NIELSEN *et al.* (1998) to detect the marker-disease association, which is given by

$$\chi^2_{\mathrm{HW}} = n\sum_{i=1}^{m} \frac{(\hat{P}_{ii|\mathrm{Affected}} - \hat{q}^2_{i|\mathrm{Affected}})^2}{\hat{q}^2_{i|\mathrm{Affected}}} + 2n\sum_{i<j} \frac{(\hat{P}_{ij|\mathrm{Affected}} - 2\hat{q}_{i|\mathrm{Affected}}\hat{q}_{j|\mathrm{Affected}})^2}{2\hat{q}_{i|\mathrm{Affected}}\hat{q}_{j|\mathrm{Affected}}},$$

where $\hat{P}_{ij|\mathrm{Affected}}$ and $\hat{q}_{i|\mathrm{Affected}}$ are the estimated frequencies

TABLE 8

Results from the NIDDM1 data

Map order	Locus	Allele frequency	No. of families	Linkage *P* value	*P* value	
					Sun *et al.*	Ours
1	SNP20	0.85	153	3.57×10^{-5}	0.0394	0.0164
2	SNP66	0.88	124	5.95×10^{-5}	0.1048	0.3536
3	SNP45	0.94	163	1.58×10^{-5}	0.0285	0.0176
4	SNP44	0.94	164	2.32×10^{-5}	0.0376	0.1462
5	SNP43	0.73	160	2.01×10^{-5}	0.0004	0.0120
6	SNP79	0.97	161	2.66×10^{-5}	0.0247	0.1798
7	SNP78	0.94	162	2.03×10^{-5}	0.0291	0.1428
8	SNP77	0.92	161	1.58×10^{-5}	0.0228	0.0688
9	SNP56	0.57	149	4.40×10^{-5}	0.0157	0.8596
10	SNP19	0.56	161	1.47×10^{-5}	0.0042	0.0016
11	SNP48	0.55	154	1.64×10^{-5}	0.0033	0.7572
12	SNP62	0.81	125	6.27×10^{-5}	0.1174	0.5374
13	SNP63	0.76	130	3.50×10^{-5}	0.0197	0.1154
14	SNP26	0.92	162	2.04×10^{-5}	0.0137	0.2748
15	SNP25	0.50	156	4.07×10^{-5}	0.0054	0.0080
16	SNP24	0.98	162	1.92×10^{-5}	0.0201	0.0874
17	SNP23	0.85	158	1.67×10^{-5}	0.0084	0.0636
18	SNP22	0.61	158	1.56×10^{-5}	0.0253	0.3362
19	SNP53	0.90	155	6.80×10^{-5}	0.0161	0.1728
20	SNP38	0.62	154	5.62×10^{-5}	0.0196	0.1198
21	SNP29	0.77	151	1.48×10^{-5}	0.0074	0.0392
22	SNP28	0.56	156	0.46×10^{-5}	0.0057	0.1868

TABLE 9

Values of observed χ^2_{HW}

j = 86,781	j = 146,317	j = 4,249,771	j = 4,169,573	j = 93,115	j = 3,116,000
34.09 (5.26×10^{-9})	8.035 (0.0046)	16.51 (0.00005)	11.728 (0.0006)	85.25 (0.0000)	31.17 (2.36×10^{-8})

of marker genotype A_iA_j and allele A_i from the observed affected individuals and m is the total number of alleles. They showed that this marker Hardy-Weinberg disequilibrium measure is proportional to the square of the disease-marker LD measure. Under the null hypothesis that there is no disease-marker LD, χ^2_{HW} is approximately distributed as a χ^2 variable with degrees of freedom $m(m-1)/2$.

After computing the value of χ^2_{HW} at each marker and their corresponding P values, we found that 13 of the markers significantly indicate strong evidence of disease-marker disequilibrium. To apply our method, we choose a set of six SNPs, and we code them as sites 1–6 for simplicity. The χ^2_{HW} values are displayed in Table 9, along with their P values in parentheses.

We see from this table that all six loci are very tightly linked to the trait. Now we use our method to identify which one of the six SNPs is the sole true cause of linkage, if any. The computed values of the conditional testing statistics and their P values are in Table 10.

From this table we see that all the P values, except that of $S_{+|31}$, are significant at the 1% level. This shows that site 3, or SNP 4249771, is most likely to be the sole cause of disease linkage for all six SNP sites.

DISCUSSION

We developed a method using the conditional LD approach to identify the true linkage-susceptible SNP in a region tightly linked to a qualitative trait, if any, using genotyping data. Simulation studies show that this method can accurately identify the true susceptibility site among a region of tightly linked loci. Application to the real data also leads to the finding of one locus, among a set of tightly linked loci, being the leading cause of linkage to the trait, while the rest of the loci are merely in tight linkage to the susceptibility locus. We illustrated the method using singleton data. This method can be applied to general pedigree data sets, in which the pedigrees are required to have homogeneous familial structure.

Our method requires only the genotype information and allele counts at each locus. It does not require phase information in diploids, which is a difficult task in contemporary sequencing and genotyping methods (Lin *et al.* 2002). Thus this method is practical to use in applications.

By forming a hypothesis that one of these sites is the sole cause and the others subordinate, we constructed testing statistics by conditioning successively on each of the sites. They can be constructed using any marker-disease LD measure based on genotype data. For illustration, our testing statistic is based on a conditional version of part of the quantity in Feder *et al.* (1996) and Nielsen *et al.* (1998), in which the relationship between marker genotype and the marker-disease LD is established. Under the true hypothesis, the testing statistic follows a mixture χ^2 distribution, with which the P values of these statistics can be obtained easily via simulation.

It is likely that the exact relevant variation goes untyped in practice; there are two possibilities for the set of SNPs under study. Some of them in the set are the susceptibility SNPs to the disease linkage, although they may not be directly disease related. Our method is designed to identify SNPs that are in tight linkage with the relevant untyped variation. When more than one SNP is identified (selected), they are not necessarily in high LD with each other, since different sources may contribute to their linkage. The other possibility is that, although showing strong disease linkage, none of them are the cause for it, or all of them are carry-ins by some untyped SNP(s) or background factors. In this case our method is expected to reject all the SNPs in the set, and a more refined scan around the region spanned by this set is suggested.

Our method is based on a set of well-chosen markers. They are chosen as a result of optimization of the corresponding model. So it is reasonable to assume the background LD to be random and negligible, and asymptotic approximation is relatively robust for such a level of noise as long as the sample size is fairly large. When some pat-

TABLE 10

Values of observed $S_{+|j1}$

j = 86,781	j = 146,317	j = 4,249,771	j = 4,169,573	j = 93,115	j = 3,116,000
5.256 (0.000)	5.415 (0.000)	1.889 (0.018)	3.080 (0.001)	5.037 (0.000)	3.626 (0.0004)

terned background is nonnegligible, one should build this effect into the model to improve the accuracy. We do not pursue this line here.

Simulation indicates our method is relatively sensitive to large deviation of HWE. In general, significant departures from HWE are not expected in practice, but if they are observed caution should be taken in applying this method. In particular, in situations in which nonrandom mating is a known confounder because of inbreeding or population structure, care should be exercised. How to modify our method to be robust against deviation from HWE will be a topic of our future research.

We thank the two anonymous reviewers, whose comments and suggestions improved the quality of the article, and Nancy Cox for providing us the NIDDM1 data. This work was supported by U.S. Public Service grant AG 16996 from the National Institutes of Health. The software used in this article is written in SAS and can be provided upon request to A.Y. or G.C. at gchen@genomecenter.howard.edu.

LITERATURE CITED

Anderson, T. W., and G. P. H. Styan, 1982 Cochran's theorem, rank additivity and tripotent matrices, pp. 1–23 in *Statistics and Probability: Essays in Honor of C. R. Rao*, edited by G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh. North-Holland, Amsterdam.

Armitage, P., 1955 Tests for linear trends in proportions and frequencies. Biometrics **11:** 375–386.

Bengtsson, B. O., and G. Thomson, 1981 Measuring the strength of associations between HLA antigens and diseases. Tissue Antigens **18:** 356–363.

Blanger, J., H. H. H. Goring, J. T. Williams, T. Dyer and L. Almasy, 2000 Quantitative trait nucleotide analysis using Bayesian model selection. Paper presented at Genetic Analysis Workshop 12, October 24–26, San Antonio, TX.

Cardon, L. R., and G. R. Abecasis, 2000 Some properties of a variance components model for fine-mapping in quantitative trait loci. Behav. Genet. **30:** 235–243.

Cox, D. R., 1972 The analysis of multivariate binary data. J. R. Stat. Soc. B **98:** 39–54.

Devlin, B., and K. Roeder, 1999 Genomic control for association studies. Biometrics **55:** 997–1004.

Feder, J. N., A. Gnirke, W. Thomas and Z. Tsuchihasi, 1996 A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat. Genet. **13:** 399–408.

Fitzmaurice, G., and N. M. Laird, 1993 A likelihood-based method for analyzing longitudinal binary responses. Biometrika **80:** 141–151.

Fulker, D. W., S. S. Cherny, P. C. Sham and J. K. Hewitt, 1999 Combined linkage and association sib-pair analysis for quantitative traits. Am. J. Hum. Genet. **64:** 259–267.

Geman, S., and D. Geman, 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Patt. Anal. Mach. Intell. **PAMI-6:** 721–741.

Good, I. J., 1969 Conditions for a quadratic form to have a chi-squared distribution. Biometrika **56:** 215–216.

Graybill, F. A., and G. Marsaglia, 1957 Idempotent matrices and quadratic forms in the general linear hypothesis. Ann. Math. Stat. **28:** 678–686.

Greenberg, D. A., 1993 Linkage analysis of "necessary" disease loci versus "susceptibility" loci. Am. J. Hum. Genet. **52:** 135–143.

Hodge, S. E., 1993 Linkage analysis versus association analysis: distinguishing between two models that explain disease marker associations. Am. J. Hum. Genet. **53:** 367–384.

Horikawa, Y., N. Oda, N. J. Cox, X. Li, M. Orho-Melander *et al.*, 2000 Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat. Genet. **26:** 163–175.

Khatri, C. G., 1980 Quadratic forms in normal variables, pp. 443–466 in *Handbook of Statistics 1*, edited by P. R. Krishnaiah. North-Holland, Amsterdam.

Khatri, C. G., 1982 A theorem on quadratic forms for normal variables, pp. 411–417 in *Statistics and Probability: Essays in Honor of C. R. Rao*, edited by G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh. North-Holland, Amsterdam.

Lazzeroni, L. C., and K. Lange, 1998 A conditional inference framework for extending the transmission/disequilibrium test. Hum. Hered. **48:** 67–81.

Lehesjoki, A.-E., M. Koskiniemi, R. Norio, S. Tirrito, P. Sistonen *et al.*, 1993 Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. Hum. Mol. Genet. **2:** 1229–1234.

Lin, S., D. J. Cutler, M. E. Zwick and A. Chakravarti, 2002 Haplotype inference in random population samples. Am. J. Hum. Genet. **71:** 1129–1137.

Nielsen, D. M., M. G. Ehm and B. S. Weir, 1998 Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am. J. Hum. Genet. **63:** 1531–1540.

Serfling, R. J., 1980 *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

Siegmund, K. D., H. Vora and W. J. Gauderman, 2001 Combined linkage and association analysis in pedigrees. Genet. Epidemiol. **21** (Suppl. 1): S358–S363.

Soria, J. M., L. Almasy, J. C. Souto, I. Tirado, M. Borell *et al.*, 2000 Linkage analysis demonstrates that the prothrombin G20210A mutation jointly influences plasma prothrombin levels and risk of thrombosis. Blood **95:** 2780–2785.

Sun, L., N. J. Cox and M. S. McPeek, 2002 A statistical method for identification of polymorphisms that explain a linkage result. Am. J. Hum. Genet. **70:** 399–411.

Tait, B. D., and L. C. Harrison, 1991 Overview: the major histocompatibility complex and insulin dependent diabetes mellitus. Baillieres Clin. Endocrinol. Metab. **5** (2): 211–228.

Thomson, G., 1991 HLA population genetics, pp. 247–260 in *The Genetics of Diabetes*, Part 1, edited by L. C. Harrison and B. D. Tait. Bailliere Tindall, London.

Valdes, A. M., and G. Thomson, 1997 Detecting disease-predisposing variants: the haplotype method. Am. J. Hum. Genet. **60:** 703–716.

Communicating editor: M. W. Feldman

APPENDIX

Proof of the proposition:

i. Let

$$\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_d) \quad \text{and} \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d).$$

Since both Σ and A are positive definite, there is an orthogonal matrix P ($PP' = P'P = I_d$) such that

$$A = P'\Gamma P \text{ (or } PAP' = \Gamma) \quad \text{and} \quad \Sigma = P'\Lambda P.$$

Let $Y = \Lambda^{-1/2}PX$ (or $X = P'\Lambda^{1/2}Y$); then Y is a normal random vector with $E(Y) = \mathbf{0}$ and

$$\begin{aligned}\mathrm{Cov}(Y) &= \Lambda^{-1/2}P\,\mathrm{Cov}(X)P'\Lambda^{-1/2} \\ &= \Lambda^{-1/2}PP'\Lambda PP'\Lambda^{-1/2} = I_d;\end{aligned}$$

i.e., $Y \sim N(\mathbf{0}, I_d)$, or its squared components $Y_1^2, \ldots, Y_d^2$ are IID χ_1^2 random variables. Now

$$\begin{aligned}X'AX &= Y'\Lambda^{1/2}PAP'\Lambda^{1/2}Y = Y'\Lambda^{1/2}\Gamma\Lambda^{1/2}Y \\ &= Y'\Gamma\Lambda Y = \gamma_1\lambda_1 Y_1^2 + \ldots + \gamma_d\lambda_d Y_d^2.\end{aligned}$$

ii. Keep notations in i, then $A^{1/2} = \Gamma^{1/2}P$. Let $Y = \Lambda^{-1/2}PX$ (or $X = P'\Lambda^{1/2}Y$). Then

$$\mathrm{Cov}(Y) = \Lambda^{-1/2}P\,\mathrm{Cov}(X)P'\Lambda^{-1/2} = I_d;$$

i.e., the Y_j's are independent standard normal random variables. Now

$$X'(A^{1/2})'\Gamma^{-1}\Lambda^{-1}X = Y'\Lambda^{1/2}PP'\Gamma^{1/2}\Gamma^{-1}\Lambda^{-1}\Gamma^{1/2}PP'\Lambda^{1/2}Y = Y'Y \sim \chi^2_d.$$

Derivation of (2): Since $X_{ik} = (X_{j|ik} : j \neq i) \sim N(\mathbf{0}, \Sigma_{ik})$ asymptotically, in the limit

$$S_{+|ik} \sim \sum_{j\neq i}\frac{X_j^2}{\lambda_j} = X'\Lambda^{-1}X,$$

where the X_j's are standard normal random variables, with $\mathrm{Cov}(X) = \Sigma_{ik}$. Since for fixed i, the $X_{j|ik}$'s are not a function of each other, neither does their distribution limit the X_j's; *i.e.*, X is a nondegenerate normal vector, and the conclusion comes from ii of the *Proposition* with $A = I_{J-1}$.

Derivation of (4): To get the asymptotic variance matrix Σ_{ik}, and hence λ, first consider the asymptotic distribution of $\sqrt{N_{ik}}(\hat{P}_{jI|ik} + \hat{P}_{jII|ik},\ \hat{q}_{j1|ik})$; then that of $\sqrt{N_{ik}}\hat{T}_{j|ik} = \sqrt{N_{ik}}(\hat{P}_{jI|ik} + \hat{P}_{jII|ik} - \hat{q}^2_{j1|ik} - \hat{q}^2_{j2|ik}) = \sqrt{N_{ik}}(\hat{P}_{jI|ik} + \hat{P}_{jII|ik} + 2\hat{q}_{j1|ik} - 2\hat{q}^2_{j1|ik} - 1)$ and that of $\sqrt{N_{ik}}\hat{T}_{ik}$ and thus that of $S_{+|ik}$ are obtained. Note that $(\hat{P}_{jI|ik} + \hat{P}_{jII|ik},\ \hat{q}_{j1|ik})$ can be written as an average of N_{ik} IID random variables, so its asymptotic normality is asserted by the central limit theorem. Let

$$g(x, y) = x + 2y - 2y^2 - 1,$$

$$\Delta g(x, y) := (\partial g/\partial x, \partial g/\partial y) = (1, 2 - 4y);$$

then under H_{ik}, $g(P_{jI|ik} + P_{jII|ik},\ q_{j1|ik}) = 0$,

$$\sqrt{N_{ik}}\hat{T}_{j|ik} = \sqrt{N_{ik}}g(\hat{P}_{jI|ik} + \hat{P}_{jII|ik},\ \hat{q}_{j1|ik}),$$

and

$$\Delta g(P_{jI|ik} + P_{jII|ik},\ q_{j1|ik}) := D_j,\ \Delta g(\hat{P}_{jI|ik} + \hat{P}_{jII|ik},\ \hat{q}_{j1|ik}) := \hat{D}_j.$$

Now using the delta method (Serfling 1980), under H_{ik}, $\sqrt{N_{ik}}(\hat{P}_{jI|ik} + \hat{P}_{jII|ik} - \hat{q}^2_{j1|ik} - \hat{q}^2_{j2|ik})$ is asymptotically $N(0,\ D_j\Sigma_{j|ik}D_j')$, where $\Sigma_{j|ik} = \mathrm{Cov}(I_{n,jI|ik} + I_{n,jII|ik},\ J_{n,j1|ik})$.

Similarly, under H_{ik}, $\sqrt{N_{ik}}T_{ik}$ is asymptotically $N(0,\ D\Sigma_{ik}D')$, and Σ_{ik} is given by

$$\Sigma_{ik} = D\Omega D',$$

where $\Omega = \mathrm{Cov}(I_{n|ik})$, and $I_{n|ik}$ is the $2(J-1)$-dimensional column vector

$$I_{n|ik} = ((I_{n,jI|ik} + I_{n,jII|ik}, J_{n,j1|ik}/2) : j \neq i),$$

and

$$D = \oplus_{j\neq i}(1,\ 2 - 4q_{j1|ik}),$$

where $\oplus$ means matrix direct summation, which results in a $(J-1)\times 2(J-1)$-dimensional matrix, and D is estimated by its empirical version $\hat{D}$ in which $q_{j1|ik}$ is replaced by $\hat{q}_{j1|ik}$. And Ω is estimated by

$$\hat{\Omega} = \frac{1}{N_{ik}-1}\sum_{n=1}^{N_{ik}} Z_{n|ik}Z'_{n|ik},$$

and

$$Z_{n|ik} = ((I_{n,jI|ik} + I_{n,jII|ik} - \hat{P}_{jI|ik} - \hat{P}_{jII|ik}, \frac{J_{n,j1|ik}}{2} - \hat{q}_{j1|ik}) : j \neq i).$$

Derivation of (5): Let $\Sigma_{A,ik}$ and $\Sigma_{U,ik}$ be the asymptotic variance matrices of $\sqrt{N_{A,ik}}(\hat{q}_{jr|A,ik})$ and $\sqrt{N_{U,ik}}(\hat{q}_{jr|U,ik})$ under their corresponding null hypothesis. Assume $N_{A,ik}/N_{ik} \to \alpha_{A,ik}$ and $N_{U,ik}/N_{ik} \to \alpha_{U,ik} = 1 - \alpha_{A,ik}$. Since $\hat{q}_{jr|A,ik}$ and $\hat{q}_{jr|U,ik}$ are independent, we have asymptotically, under their corresponding null hypothesis,

$$\sqrt{N_{ik}}(\hat{q}_{jr|A,ik},\ \hat{q}_{jr|U,ik})' \xrightarrow{d} N(\mathbf{0},\ \Omega_{jr|ik}),$$

where

$$\Omega_{ik} = \begin{pmatrix}\alpha^{-1}_{A,ik}\Sigma_{A,ik} & 0\\ 0 & \alpha^{-1}_{U,ik}\Sigma_{U,ik}\end{pmatrix}.$$

Let $g(x, y) = (x - y)/(1 - y)$; then $\Delta g(x, y) := (\partial g/\partial x, \partial g/\partial y) = (1/(1-y),\ (x-1)/(1-y)^2)$. Under H_{ik}, $R(jr|ik) = 0$, thus by the delta method,

$$\sqrt{N_{ik}}\hat{R}(jr|ik) \xrightarrow{d} N(\mathbf{0},\ \Sigma_{jr|ik}),$$

where

$$\Sigma_{jr|ik} = \Delta g(q_{jr|A,ik},\ q_{jr|U,ik})\Omega_{jr|ik}\Delta g(q_{jr|A,ik},\ q_{jr|U,ik})' = \alpha^{-1}_{A,ik}q^2_{jr|A,ik}\Sigma_{A,ik} + \alpha^{-1}_{U,ik}q^2_{jr|U,ik}\Sigma_{U,ik}.$$

Similarly,

$$\sqrt{N_{ik}}\hat{R}_{ik} \xrightarrow{d} N(\mathbf{0},\ \Sigma_{ik}),$$

for some Σ_{ik}, where

$$\hat{R}_{ik} = (\hat{R}(j1|ik) : j \neq i).$$

Let

$$J^A_{n|ik} = (J^A_{n,j1|ik} : j \neq i),\quad J^U_{n|ik} = (J^U_{n,j1|ik} : j \neq i),$$
$$J_{n|ik} = (J^A_{n|ik}, J^U_{n|ik}).$$

Let Ω_A, Ω_U, and Ω be the asymptotic variance matrices for $\sqrt{N_{A,ik}}J^A_{n|ik}$, $\sqrt{N_{U,ik}}J^U_{n|ik}$, and $\sqrt{N_{ik}}J_{n|ik}$. Let

$$D = \oplus_{j\neq i}\left(\frac{1}{1 - q_{j1|A,ik}}, \frac{q_{j1|A,ik} - 1}{(1 - q_{j1|U,ik})^2}\right).$$

The same way as before,

$$\Sigma_{ik} = D\Omega D'.$$

Derivation of (6): Let $\hat{R}_i = (\hat{R}_{i1}, \hat{R}_{i2})$. Under H_i, asymptotically $\sqrt{N_{ik}}\hat{R}_i \sim N(\mathbf{0},\ \Sigma_i)$ for some matrix Σ_i. Let

$$J^A_{n,r|ik} = (J^A_{n,jr|ik} : j \neq i)\quad (r = 1, 2)$$
$$J^A_{n|ik} = (J^A_{n,1|ik}, J^A_{n,2|ik}),\ J^U_{n|ik} = (J^U_{n,1|ik}, J^U_{n,2|ik}),$$
$$J_{n|ik} = (J^A_{n|ik}, J^U_{n|ik}).$$

Let $N_i = N_{i1} + N_{i2}$, $N_{A,i} = N_{A,i1} + N_{A,i2}$, $N_{U,i} = N_{U,i1} + N_{U,i2}$,

$\alpha_{A,i} = N_{A,i}/N_i$, $\alpha_{U,i} = N_{U,i}/N_i = 1 - \alpha_{A,i}$. Let Ω_A, Ω_U, and Ω be the asymptotic variance matrices for $\sqrt{N_{A,i}}J^A_{n|i}$, $\sqrt{N_{U,i}}J^U_{n|i}$, and $\sqrt{N_i}J_{n|i}$. Let

$$D = \oplus_{j\neq i,k=1,2}\left(\frac{1}{1 - q_{j1|A,ik}}, \frac{q_{j1|A,ik} - 1}{(1 - q_{j1|U,ik})^2}\right).$$

Then similarly as the derivation of (4) we have

$$\Sigma_i = D\Omega D'.$$

Derivation of (7): We need only to derive, under H_{ik}, the asymptotic distribution of $\sqrt{M_{ik}}\hat{T}_{ik}$. We first derive that of

$$\sqrt{M_{ik}}\hat{T}_{j|ik} = \sqrt{M_{ik}}\left(\sum_{l=1}^{L}\frac{M^{(l)}_{ik}}{M_{ik}}(\hat{P}^{(l)}_{jI|ik} + \hat{P}^{(l)}_{jII|ik}) - \left(\sum_{l=1}^{L}\frac{M^{(l)}_{ik}}{M_{ik}}\hat{q}^{(l)}_{j1|ik}\right)^2 - \left(\sum_{l=1}^{L}\frac{M^{(l)}_{ik}}{M_{ik}}\hat{q}^{(l)}_{j2|ik}\right)^2\right)$$

for each j. Again, we first get the asymptotic distribution of

$$\sqrt{M_{ik}}\sum_{l=1}^{L}\frac{M^{(l)}_{ik}}{M_{ik}}(\hat{P}^{(l)}_{jI|ik} + \hat{P}^{(l)}_{jII|ik}, \hat{q}^{(l)}_{j1|ik}). \quad \text{(A1)}$$

The summands above are independent of each other, and recall $\alpha^{(l)}_{ik} = \lim M^{(l)}_{ik}/M_{ik}$. Since $\sqrt{M^{(l)}_{ik}}(\hat{P}^{(l)}_{jI|ik} + \hat{P}^{(l)}_{jII|ik}, \hat{q}^{(l)}_{j1|ik})$ is asymptotically $N(\mathbf{0}, \Omega^{(l)}_j)$, with

$$\Omega^{(l)}_j = \mathrm{Cov}(I^{(l)}_{n,jI|ik} + I^{(l)}_{n,jII|ik}, J^{(l)}_{n,j1|ik}/2),$$

by Slutsky's theorem, (A1) is asymptotically $N(\mathbf{0}, \Omega_j)$ with

$$\Omega_j = \sum_{l=1}^{L}\alpha^{(l)}_{ik}\Omega^{(l)}_j.$$

Let $g(x, y)$ be the same as in the derivation of (4), and

$$D_j = \Delta g\left(\sum_{l=1}^{L}\alpha^{(l)}_{ik}(P^{(l)}_{jI|ik} + P^{(l)}_{jII|ik}, q^{(l)}_{j1|ik})\right) = (1, 2 - 4q_{j1|ik}).$$

Under H_{ik}, $g(\Sigma_{l=1}^{L}\alpha^{(l)}_{ik}(P^{(l)}_{jI|ik} + P^{(l)}_{jII|ik}, q^{(l)}_{j1|ik})) = 0$, and

$$\sqrt{M_{ik}}\hat{T}_{j|ik} = \sqrt{M_{ik}}g\left(\sum_{l=1}^{L}\frac{M^{(l)}_{ik}}{M_{ik}}(\hat{P}^{(l)}_{jI|ik} + \hat{P}^{(l)}_{jII|ik}, \hat{q}^{(l)}_{j1|ik}1)\right).$$

So $\sqrt{M_{ik}}\hat{T}_{j|ik}$ is asymptotically normal with zero mean vector and variance matrix

$$\Sigma_{j|ik} = D_j\Omega_j D'_j = D_j\sum_{l=1}^{L}\alpha^{(l)}_{ik}\Omega^{(l)}_j D'_j.$$

Now the final conclusion follows the same way as in the derivation of (4).