DOI: 10.1534/genetics.105.045781

Improvement of Mapping Accuracy by Unifying Linkage and Association Analysis

Xiang-Yang Lou,* Jennie Z. Ma,† Mark C. K. Yang,‡ Jun Zhu,§ Peng-Yuan Liu, Hong-Wen Deng,** Robert C. Elston†† and Ming D. Li*,[1]**

**Department of Psychiatric Medicine, University of Virginia, Charlottesville, Virginia 22911,* †*Department of Psychiatry, University of Texas Health Science Center, San Antonio, Texas 78229,* ‡*Department of Statistics, University of Florida, Gainesville, Florida 32611,* §*Department of Agronomy, Zhejiang University, Hangzhou, Zhejiang 310029, People's Republic of China,* ***Osteoporosis Research Center, Creighton University Medical Center, Omaha, Nebraska 68131 and* ††*Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44109*

Manuscript received May 17, 2005
Accepted for publication September 14, 2005

ABSTRACT

It is well known that pedigree/family data record information on the coexistence in founder haplotypes of alleles at nearby loci and the cotransmission from parent to offspring that reveal different, but complementary, profiles of the genetic architecture. Either conventional linkage analysis that assumes linkage equilibrium or family-based association tests (FBATs) capture only partial information, leading to inefficiency. For example, FBATs will fail to detect even very tight linkage in the case where no allelic association exists, while a violation of the assumption of linkage equilibrium will result in biased estimation and reduced efficiency in linkage mapping. In this article, by using a data augmentation technique and the EM algorithm, we propose a likelihood-based approach that embeds both linkage and association analyses into a unified framework for general pedigree data. Relative to either linkage or association analysis, the proposed approach is expected to have greater estimation accuracy and power. Monte Carlo simulations support our theoretical expectations and demonstrate that our new methodology: (1) is more powerful than either FBATs or classic linkage analysis; (2) can unbiasedly estimate genetic parameters regardless of whether association exists, thus remedying the bias and less precision of traditional linkage analysis in the presence of association; and (3) is capable of identifying tight linkage alone. The new approach also holds the theoretical advantage that it can extract statistical information to the maximum extent and thereby improve mapping accuracy and power because it integrates multilocus population-based association study and pedigree-based linkage analysis into a coherent framework. Furthermore, our method is numerically stable and computationally efficient, as compared to existing parametric methods that use the simplex algorithm or Newton-type methods to maximize high-order multidimensional likelihood functions, and also offers the computation of Fisher's information matrix. Finally, we apply our methodology to a genetic study on bone mineral density (BMD) for the vitamin D receptor (VDR) gene and find that VDR is significantly linked to BMD at the one-third region of the wrist.

TWO approaches are commonly used in pedigree- or family-based gene mapping, *i.e.*, linkage analysis (*e.g.*, Elston and Stewart 1971; Haseman and Elston 1972; Ott 1974; Lander and Green 1987; Risch 1990; Ward 1993; Amos 1994; Kruglyak and Lander 1995; O'Connell and Weeks 1995; Kruglyak *et al.* 1996; Gudbjartsson *et al.* 2000; Abecasis *et al.* 2002) and family-based association tests (FBATs) (*e.g.*, Falk and Rubinstein 1987; Spielman *et al.* 1993; Lazzeroni and Lange 1998; Laird *et al.* 2000; Rabinowitz and Laird 2000). Linkage analysis focuses on gene cosegregation that can be characterized by inheritance vectors or gene concordance between related individuals (identical-by-descent, IBD, or identical-in-state, IIS) at each locus, while association tests (which, when due to linkage, are tests of gametic association, also called linkage disequilibrium, LD) directly utilize allele status and linkage phase that record historic events. Pedigree data contain both these components of information that give rise to complementary profiles of the genetic architecture. Either linkage or association analysis alone, however, can capitalize only on the genetic information from one of these components and fails to grasp the whole picture, thereby leading to a loss in mapping accuracy and statistical power.

To illustrate the limitations of applying either a linkage or association approach alone, let us consider the affected sib pair design used in Risch (1990) and Risch and Merikangas (1996). First, traditional linkage analysis will give a biased result in the presence of population association. To simplify our exposition, assume there are a diallelic disease locus $\mathcal{Q}$ with alleles Q and q and a codominant marker locus $\mathcal{A}$ with alleles A and a. Alleles Q and A have the same frequency and are in perfect association, and let $p_Q = p_A = p_{AQ} = p$. Table 1 lists the assumed probabilities (under no association),

[1]*Corresponding author:* 1670 Discovery Dr., Ste. 110, Charlottesville, VA 22911. E-mail: ml2km@virginia.edu

TABLE 1

Probabilities and Risch's (1990) LOD scores in affected sib-pairs designs for a marker unlinked to, but perfectly associated with, a recessive disease gene

Sib configuration	Assumed probability (when $p = 0.5$)[a]	True probability (when $p = 0.5$)	LOD score[b]
AA, AA	$\alpha_2 p^2 + \alpha_1 p^3 + \alpha_0 p^4 (\frac{9}{64})$	$\frac{(1+7p)^2}{16(1+p)^2}\left(\frac{9}{16}\right)$	$\log\frac{\hat{z}_2 p^2 + \hat{z}_1 p^3 + \hat{z}_0 p^4}{\alpha_2 p^2 + \alpha_1 p^3 + \alpha_0 p^4}$
AA, Aa	$\alpha_1[2p^2(1-p)] + \alpha_0[4p^3(1-p)] (\frac{3}{16})$	$\frac{(1-p)(1+7p)}{4(1+p)^2}\left(\frac{1}{4}\right)$	$\log\frac{\hat{z}_1[2p^2(1-p)] + \hat{z}_0[4p^3(1-p)]}{\alpha_1[2p^2(1-p)] + \alpha_0[4p^3(1-p)]}$
AA, aa	$\alpha_0[2p^2(1-p)^2] (\frac{1}{32})$	$\frac{(1-p)^2}{8(1+p)^2}\left(\frac{1}{72}\right)$	$\log\frac{\hat{z}_0}{\alpha_0}$
Aa, Aa	$\alpha_2[2p(1-p)] + \alpha_1[p(1-p)] + \alpha_0[4p^2(1-p)^2] (\frac{5}{16})$	$\frac{(1-p)(1+3p)}{4(1+p)^2}\left(\frac{5}{36}\right)$	$\log\frac{\hat{z}_2[2p(1-p)] + \hat{z}_1[p(1-p)] + \hat{z}_0[4p^2(1-p)^2]}{\alpha_2[2p(1-p)] + \alpha_1[p(1-p)] + \alpha_0[4p^2(1-p)^2]}$
Aa, aa	$\alpha_1[2p(1-p)^2] + \alpha_0[4p(1-p)^3] (\frac{3}{16})$	$\frac{(1-p)^2}{4(1+p)^2}\left(\frac{1}{36}\right)$	$\log\frac{\hat{z}_1[2p(1-p)^2] + \hat{z}_0[4p(1-p)^3]}{\alpha_1[2p(1-p)^2] + \alpha_0[4p(1-p)^3]}$
aa, aa	$\alpha_2(1-p)^2 + \alpha_1(1-p)^3 + \alpha_0(1-p)^4 (\frac{9}{64})$	$\frac{(1-p)^2}{16(1+p)^2}\left(\frac{1}{144}\right)$	$\log\frac{\hat{z}_2(1-p)^2 + \hat{z}_1(1-p)^3 + \hat{z}_0(1-p)^4}{\alpha_2(1-p)^2 + \alpha_1(1-p)^3 + \alpha_0(1-p)^4}$

[a] α_i ($i = 0, 1, 2$) is the prior probability that two siblings share i alleles IBD, $\alpha_2 = \alpha_0 = 0.25$, $\alpha_1 = 0.5$, respectively.
[b] $\hat{z}_i$ ($i = 0, 1, 2$) is the estimated posterior probability that two affected siblings share i alleles IBD.

the true probabilities, and Risch's (1990) LOD scores of all six possible sib configurations in the case where marker $\mathcal{A}$ is unlinked to a recessively inherited disease gene $\mathcal{Q}$. Using Risch's (1990) EM iterative Equation 4, we can obtain the maximum-likelihood estimates (MLEs) of the posterior probabilities that the affected sib pairs share i marker alleles IBD ($i = 0, 1, 2$). To illustrate the result, we take p to be a specific value, say $p = 0.5$, then we have $\hat{z}_2 = 0.444$, $\hat{z}_2 = 0.487$, and $\hat{z}_0 = 0.069$, respectively, and the expected LOD (ELOD) = 0.384. These values deviate substantially from the true IBD sharing scores of 0.25, 0.5, and 0.25, respectively, and exhibit a spuriously excessive allele sharing. This suggests that a false-positive result can occur in allele-sharing analysis. We further demonstrate that, generally, the assumed likelihood is a monotonically decreasing function of the recombination fraction θ for $\theta \in [0, 0.5]$ (see the appendix). This means that, if the true recombination fraction $\theta_0 \neq 0$, we may still obtain an estimate of zero.

Second, neglecting to take account of information on association may cause loss of statistical power. As pointed out by Risch and Merikangas (1996), the allele-sharing method is much less powerful than the transmission/disequilibrium test (TDT) method in the cases they considered, *i.e.*, when there is no recombination and the alleles at the two loci are perfectly associated. This arises because the linkage statistic, the mean allele sharing, fails to consider the *allele-specific* IBD sharing. Actually, allele *A* (increasing disease risk) contributes more allele sharing to the statistic, whereas allele *a* contributes less, so that the overall mean allele sharing is diluted. Our simulations of model-based linkage-only analysis support this theoretical argument, *i.e.*, the plausible bias and the reduced power (see simulation studies).

Because they fail to incorporate information on linkage, FBATs are inherently conservative, and so they cannot detect linkage even when two or more siblings are available, unless there is also population association. The conclusion by Risch and Merikangas (1996) was drawn from the ideal circumstance where the marker is the disease gene itself. In such a situation, FBATs reach their maximum potential power. In practice, however, it may not be true that a marker happens to have the same variant frequencies as, and be perfectly associated with, the disease gene of interest, even for fine mapping, as there are always many polymorphic SNPs within a gene whereas only a few may be responsible for the change of its function. Both theoretical and empirical studies (*e.g.*, Kruglyak 1999; Hinds *et al.* 2005) have shown that the founder LD within a small region has usually been largely disrupted by various population forces, such as recombination, gene conversion, and/or mutation accumulated over time, so that high-LD regions with little genetic shuffling, termed *haplotype blocks*, span only a very short distance, implying that strong LD is not inevitable with tightly linked loci. HapMap studies also indicate that the frequencies of variants change from one SNP to another largely within a block (International HapMap Consortium 2003). In practical application, FBATs can therefore lose their theoretical power even with closely linked loci, owing to the violation of such an ideal assumption. Furthermore, association may extend over a great distance, even to nonsyntenic loci because of factors other than linkage, such as population subdivision and admixture, population bottlenecks, mutation, gene conversion, meiotic drive, sampling or ascertainment bias, nonrandom mating, and coancestry. Caution is also required in that a positive result

from an FBAT does not necessarily imply the presence of tight linkage; *i.e.*, an FBAT alone cannot distinguish strong association and loose linkage from weak association and tight linkage (Elston 1998; Whittaker *et al.* 2000).

Therefore, it is of great interest to remedy the above limitations. A judicious way is to take both these pieces of information into consideration in gene mapping. Such an idea was conceived in earlier literature (*e.g.*, MacLean *et al.* 1984) and adopted in some computer software such as LINKAGE (Lathrop and Lalouel 1984; Lathrop *et al.* 1985). Unfortunately, the bonus from joint mapping was not recognized, so this remarkable idea has been buried for several years (Xiong and Jin 2000). Recently, Zhao *et al.* (1998) proposed a semiparametric method for a combined linkage and linkage disequilibrium analysis. Xiong and Jin (2000) advocated a likelihood-based parametric method for joint analysis with nuclear family data. Cantor *et al.* (2005) further extended Xiong and Jin's (2000) method for general pedigrees. Li *et al.* (2005) suggested an approach that identifies associated and potentially causal SNPs through joint modeling of linkage and association. Parallel to parametric ones, variance components (*e.g.*, Allison *et al.* 1999; Fulker *et al.* 1999; Abecasis *et al.* 2000) and nonparametric (Huang and Jiang 1999; Wicks 2000; Wicks and Wilson 2000; Lazzeroni 2002) methods have also been developed. However, those methods work mostly for specific data structures and types such as affected sib pairs, nuclear families, and categorical traits and/or can provide a solution only for specific problems such as single-point analysis. The bonus of combined mapping has also not been thoroughly explored. By invoking a data augmentation technique and the EM algorithm, we have evolved a general likelihood-based statistical framework for integrating linkage and association analyses (Lou *et al.* 2005). In the present article, we further extend this model-based approach for general pedigrees. This approach allows us to simultaneously perform segregation, linkage, and association analyses, *i.e.*, to estimate penetrance functions, genetic distances, and association parameters, as well as to carry out the corresponding hypothesis tests within a unified framework. More appealingly, it adds several unique strengths to existing parametric methods (*e.g.*, Xiong and Jin 2000; Cantor *et al.* 2005; Li *et al.* 2005). First, this framework is conceptually straightforward, flexible, easy to generalize, and also comprehensive, so that it covers a wide range of cases with multiple loci and/or multiple alleles. Multilocus mapping and epistatic QTL mapping can be implemented as well under the same concept. Second, our new approach is computationally efficient and powerful. We formulated the closed-form solutions for MLEs implemented with EM iteration and thus avoid the computational difficulty of high-order multidimensional searches, leading to less computational time per iteration and quick convergence. Third, due to the advantage of the EM algorithm over the simplex algorithm and Newton-type methods in the context of a mapping study, as pointed out by some authors (*e.g.*, Lander and Green 1987), our new approach is numerically stable, as compared with existing methods. In our experience, a wide range of initial values appears to give good convergence. Finally, we offer the computation of Fisher's information matrix and hence can provide the estimation precision of MLEs. Although this article emphasizes a demonstration of the improvement in mapping accuracy using a two-locus model, *i.e.*, one marker and one trait gene, we use an interval mapping model to describe our new approach in the model and method section for readers to have a clearer picture about it. After presenting the theory, we use simulation studies to compare the power of an FBAT, of the pure linkage method, and of our new approach and the estimation precision of the latter two. An application to the genetic study of bone mineral density (BMD) is used to demonstrate this new methodology. Finally, we discuss some relevant issues to provide further insights into this approach.

MODEL AND METHOD

Here we use a three-diallelic-locus model to illustrate the approach. Suppose there are three loci, one trait gene or QTL, $\mathcal{Q}$, bracketed by a pair of flanking markers, $\mathcal{A}$ and $\mathcal{B}$, respectively. Let *A*, *a*, *Q*, *q*, *B*, and *b* be the alleles at the three loci, respectively. All the alleles together form eight *haplotypes*, *AQB*, *AQb*, *AqB*, *Aqb*, *aQB*, *aQb*, *aqB*, and *aqb*. These haplotypes unite to generate a total of 36 *diplotypes*, *AQB/AQB*, *AQB/AQb*, ..., and *aqb/aqb*, where the "/" denotes the separation of the maternally and paternally derived gametes. The 36 diplotypes are collapsed into 27 *zygote genotypes*, each with an identical allelic combination at all the loci, and further, into 9 marker genotypes and 3 QTL genotypes. Owing to the fact that genotypes are *conflated data* that ignore the linkage phases of diplotypes, some of the genotypes consist of >1 diplotype. For example, all 4 diplotypes *AQB/aqb*, *AQb/aqB*, *AqB/aQb*, and *Aqb/aQB* exhibit the same genotype, *AaQqBb*. To express the relationship between diplotypes and genotypes, we denote by $\mathcal{G}(\cdot)$, $\mathcal{G}_{\mathbf{m}}(\cdot)$, and $\mathcal{G}_{\mathbf{q}}(\cdot)$ the many–one mapping operators taking the genotypes at all loci, the marker loci and the QTL, of a diplotype in parentheses, respectively. Thus, $\mathcal{G}(AQB/aqb)=\mathcal{G}(AQb/aqB)=\mathcal{G}(AqB/aQb)=\mathcal{G}(Aqb/aQB)=AaQqBb$, $\mathcal{G}_{\mathbf{m}}(AQB/aqb)=\mathcal{G}_{\mathbf{m}}(AQb/aqB)=\mathcal{G}_{\mathbf{m}}(AqB/aQb)=\mathcal{G}_{\mathbf{m}}(Aqb/aQB)=AaBb$, and $\mathcal{G}_{\mathbf{q}}(AQB/aqb)=\mathcal{G}_{\mathbf{q}}(AQb/aqB)=\mathcal{G}_{\mathbf{q}}(AqB/aQb)=\mathcal{G}_{\mathbf{q}}(Aqb/aQB)=Qq$.

We use $p_{AQB}, p_{AQb}, \ldots, p_{aqb}$ and $P_{AQB/AQB}, P_{AQB/AQb}, \ldots, P_{aqb/aqb}$ to denote the frequencies of the haplotypes *AQB*, *AQb*, ..., *aqb* and diplotypes *AQB/AQB*, *AQB/AQb*, ..., *aqb/aqb*, respectively, in the population studied. If the population is at Hardy–Weinberg equilibrium, we have

$$P_{AQB/AQB} = p_{AQB}^2,\ P_{AQB/AQb} = 2p_{AQB}p_{AQb},\ \ldots$$

The frequencies of the haplotypes can be decomposed into different components determined by the allele frequencies at each locus and LD coefficients of different orders; *e.g.*,

$$p_{AQB} = p_A p_B p_Q + p_A D_{BQ} + p_B D_{AQ} + p_Q D_{AB} + D_{AQB},$$

where p_A, p_B, and p_Q are the frequencies of alleles *A*, *B*, and *Q*, respectively and D_{AQ}, D_{BQ}, D_{AB}, and D_{AQB} are the LD coefficients, respectively. Reversely, the frequencies of alleles and LD coefficients can also be represented by the frequencies of haplotypes; *e.g.*, $p_A = p_{AQB} + p_{AQb} + p_{AqB} + p_{Aqb}, \ldots, D_{AB} = p_{AB} - p_A p_B, \ldots$, and $D_{AQB} = p_{AQB} - p_A D_{BQ} - p_B D_{AQ} - p_Q D_{AB} - p_A p_B p_Q$. For a more general expression with an arbitrary number of alleles and/or loci, see one of our recent communications (Lou *et al.* 2003) for details.

Crossing over between a pair of contiguous loci may take place during meiosis. Either recombination (*R*) or nonrecombination (*N*) between each of the pairs of adjacent loci (*i.e.*, $\mathcal{A}$ and $\mathcal{Q}$, $\mathcal{B}$ and $\mathcal{Q}$) will give rise to four recombination configurations described by *NN*, *NR*, *RN*, or *RR*. The frequency of a new haplotype is a function of the recombination fraction(s) associated with its recombination configuration(s). For simplicity, we here ignore crossover interference during gametogenesis. Let θ_{AQ} and θ_{BQ} be the recombination fractions between loci $\mathcal{A}$ and $\mathcal{Q}$ and between $\mathcal{B}$ and $\mathcal{Q}$, respectively. The frequencies of these four configurations can be expressed in terms of θ_{AQ} and θ_{BQ}, *i.e.*, $(1-\theta_{AQ})(1-\theta_{BQ})$, $(1-\theta_{AQ})\theta_{BQ}$, $\theta_{AQ}(1-\theta_{BQ})$, or $\theta_{AQ}\theta_{BQ}$ corresponding to *NN*, *NR*, *RN*, or *RR*, respectively. Furthermore, the conditional probability of a zygote randomly formed by the haplotypes generated from a pair of parents is a product of the frequencies of paternally and maternally original haplotypes.

For any complex trait, either continuous or discrete, there is no one–one correspondence between genotype and phenotype. The conditional probability of observing a phenotype given a specified genotype, termed the *penetrance function*, is thus used to characterize the relationship between genotype and phenotype. Because the phenotype is genetically determined by the genotypes at locus $\mathcal{Q}$, the penetrance function, given diplotype $\mathcal{D}$, can be expressed as

$$f(y|\mathcal{D}) = f(y|\mathcal{G}_{\mathbf{q}}(\mathcal{D})) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-\mu_{\mathcal{G}_{\mathbf{q}}(\mathcal{D})})^2}{2\sigma^2}\right],$$

for a continuous phenotype in which it is typically assumed that the distribution within each subpopulation defined by genotype is normal, where $\mu_{\mathcal{G}_{\mathbf{q}}(\mathcal{D})}$ is the genotypic mean of QTL genotype $\mathcal{G}_{\mathbf{q}}(\mathcal{D})(= QQ, Qq, \text{or } qq)$ and σ^2 is the residual variance. For a categorical trait the penetrance $f(y|\mathcal{G}_{\mathbf{q}}(\mathcal{D}))$ is defined as the probability that individuals with genotype $\mathcal{G}_{\mathbf{q}}(\mathcal{D})$ manifest phenotype *y*. We may specify different penetrance functions to mothers, fathers, and children on the basis of the inheritance pattern of the trait under investigation. To make this presentation terser, here we assume the same penetrance for the parental and offspring generations. However, it is not difficult to recast the methodology to be applicable to the case with different penetrance functions. Mendelian trait(s) and marker(s) can be viewed as specific examples with full penetrance. Then the methodology developed hereinafter is also applicable to their analysis.

In a gene-mapping study aimed at estimating parameters of penetrance, association, and position (usually measured by the recombination fractions), a major challenge is that *latent data* exist, also referred to as *missing data*, that cannot be directly observed, such as disease genotype, diplotype, and recombination configuration. We hypothesize the observed data, *i.e.*, marker genotypes and phenotypes, together with the latent data, *i.e.*, diplotypes and recombination configurations, as *complete data*, also termed *augmented data*. Correspondingly, the observed data alone are called *incomplete data*. The observed data can be viewed as mixtures of complete data and then we can use a mixture model to tackle the issue of parameter estimation.

The complete data likelihood: Denote marker, diplotype/haplotype, recombination configuration, and phenotype data by $\mathbf{M}$, $\mathcal{D}/\mathcal{H}$, $\mathcal{R}$, and $\mathbf{y}$, respectively. Observed marker and phenotypic data are in boldface type while the missing data for parent and child diplotypes and child recombination configurations are in script type. We first use nuclear family data, in which there is no phenotypic covariance between parents and children, to demonstrate parameter estimation within a unified framework of interval mapping and LD mapping, and then extend the method to general pedigree data.

With *N* unrelated nuclear families randomly drawn from a general population, the overall likelihood is the product of individual family likelihoods, denoted $L_1, L_2, \ldots, L_N$. Let us present an example to demonstrate how to build the likelihood function. In the example, family *i* consists of a mother with diplotype *AQB/AQB* ($\mathcal{D}_i^{\mathrm{m}}$) and phenotype y_i^{m}, a father with *AQB/Aqb* ($\mathcal{D}_i^{\mathrm{f}}$) and y_i^{f}, and two children with diplotypes and recombination configurations *AQB/AQB* and *NN/RN* ($\mathcal{D}_{i1}^{\mathrm{o}}$, $\mathcal{R}_{i1}$) and *AQB/AQb* and *NN/NR* ($\mathcal{D}_{i2}^{\mathrm{o}}$, $\mathcal{R}_{i2}$), respectively, and phenotypes y_{i1}^{o} and y_{i2}^{o}, respectively. The likelihood can be expressed by a three-level hierarchical model,

$$\begin{aligned} L_i &= L(y_i^{\mathrm{m}}, y_i^{\mathrm{f}}, y_{i1}^{\mathrm{o}}, y_{i2}^{\mathrm{o}}, \mathcal{D}_i^{\mathrm{m}}, \mathcal{D}_i^{\mathrm{f}}, \mathcal{D}_{i1}^{\mathrm{o}}, \mathcal{R}_{i1}, \mathcal{D}_{i2}^{\mathrm{o}}, \mathcal{R}_{i2}|\mathbf{\Omega}) \\ &\propto \Pr(\mathcal{D}_i^{\mathrm{m}})\Pr(\mathcal{D}_i^{\mathrm{f}}) f(y_i^{\mathrm{m}}|\mathcal{D}_i^{\mathrm{m}}) f(y_i^{\mathrm{f}}|\mathcal{D}_i^{\mathrm{f}}) \\ &\quad \times \prod_{j=1}^{2}\left[\Pr(\mathcal{D}_{ij}^{\mathrm{o}}, \mathcal{R}_{ij}|\mathcal{D}_i^{\mathrm{m}}, \mathcal{D}_i^{\mathrm{f}}) f(y_{ij}^{\mathrm{o}}|\mathcal{D}_{ij}^{\mathrm{o}})\right] \\ &\propto p_{AQB}^3 p_{Aqb}\theta_{AQ}(1-\theta_{AQ})^3\theta_{BQ}(1-\theta_{BQ})^3 \\ &\quad \times f(y_i^{\mathrm{m}}|QQ) f(y_i^{\mathrm{f}}|Qq) f(y_{i1}^{\mathrm{o}}|QQ) f(y_{i2}^{\mathrm{o}}|QQ), \end{aligned}$$

where $\mathbf{\Omega}$ is the vector of unknown parameters containing three subsets of population genetic parameters (haplotype frequencies, $\mathbf{\Omega}_P$), penetrance parameters

(*e.g.*, genotypic values and the residual variance, $\mathbf{\Omega}_Q$), and position parameters (recombination fractions, $\mathbf{\Omega}_R$), related to the parental diplotype distribution, the phenotype density functions, and $\Pr(\mathcal{D}_{ij}^{\mathrm{o}}, \mathcal{R}_{ij}|\mathcal{D}_i^{\mathrm{m}}, \mathcal{D}_i^{\mathrm{f}})$, respectively. $\Pr(\mathcal{D}_{ij}^{\mathrm{o}}, \mathcal{R}_{ij}|\mathcal{D}_i^{\mathrm{m}}, \mathcal{D}_i^{\mathrm{f}})$ represents the conditional probability of child j of family i having diplotype $\mathcal{D}_{ij}^{\mathrm{o}}$ and recombination configuration $\mathcal{R}_{ij}$ given parental diplotypes $\mathcal{D}_i^{\mathrm{m}}$ and $\mathcal{D}_i^{\mathrm{f}}$. The overall likelihood can be represented as

$$
\begin{aligned}
&L(\mathbf{y}^{\mathrm{m}}, \mathbf{y}^{\mathrm{f}}, \mathbf{y}^{\mathrm{o}}, \mathcal{D}^{\mathrm{m}}, \mathcal{D}^{\mathrm{f}}, \mathcal{D}^{\mathrm{o}}, \mathcal{R}|\mathbf{\Omega}) \\
&\quad = \prod_{i=1}^{N} L_i \propto \prod_{i=1}^{N} \left\{ \begin{array}{l} \Pr(\mathcal{D}_i^{\mathrm{m}})\Pr(\mathcal{D}_i^{\mathrm{f}}) f(y_i^{\mathrm{m}}|\mathcal{D}_i^{\mathrm{m}}) f(y_i^{\mathrm{f}}|\mathcal{D}_i^{\mathrm{f}}) \\ \quad \times \prod_{j=1}^{N_i} \left[\Pr(\mathcal{D}_{ij}^{\mathrm{o}}, \mathcal{R}_{ij}|\mathcal{D}_i^{\mathrm{m}}, \mathcal{D}_i^{\mathrm{f}}) f(y_{ij}^{\mathrm{o}}|\mathcal{D}_{ij}^{\mathrm{o}}) \right] \end{array} \right\} \\
&\quad \propto p_{AQB}^{n_{AQB}} p_{AQb}^{n_{AQb}} \cdots p_{aqb}^{n_{aqb}} \theta_{AQ}^{n_{\theta_{AQ}}} (1-\theta_{AQ})^{n_{\bar{\theta}_{AQ}}} \theta_{BQ}^{n_{\theta_{BQ}}} (1-\theta_{BQ})^{n_{\bar{\theta}_{BQ}}} \\
&\qquad \times \prod_{i=1}^{N} \left[f(y_i^{\mathrm{m}}|\mathcal{D}_i^{\mathrm{m}}) f(y_i^{\mathrm{f}}|\mathcal{D}_i^{\mathrm{f}}) \prod_{j=1}^{N_i} f(y_{ij}^{\mathrm{o}}|\mathcal{D}_{ij}^{\mathrm{o}}) \right], \qquad (1)
\end{aligned}
$$

where the $\mathbf{y}$'s are the phenotypic vectors; $\mathcal{D}$'s are the diplotype vectors; $\mathcal{R}$ is the recombination configuration for the children; N_i is the number of children within family i; $n_{AQB}, n_{AQb}, \ldots, n_{aqb}$ are the numbers of haplotypes AQB, AQb, $\ldots$, aqb appearing in parental diplotypes, respectively; $n_{\theta_{AQ}}$ and $n_{\bar{\theta}_{AQ}}$ are the numbers of recombinants and nonrecombinants between loci $\mathcal{A}$ and $\mathcal{Q}$ existing in the recombination configurations, respectively; and $n_{\theta_{BQ}}$ and $n_{\bar{\theta}_{BQ}}$ are those between $\mathcal{B}$ and $\mathcal{Q}$, respectively.

In many cases, information is partial because of experimental errors, financial limitations, or other practical constraints, as often occurs in studies of late-onset diseases such as Alzheimer's disease where parents are unavailable. Since missing phenotypic observations can be treated by simply setting the corresponding $f(y|D)$'s equal to 1 wherever they occur in the above likelihood, Equation 1 automatically covers the likelihoods of family data with missing phenotypes like TDT-type data. For data with missing diplotypes such as sibship data, instead of Equation 1 we can use a form of mixture model summing over all plausible diplotypes and/or recombination configurations compatible with the available data to represent such likelihoods and so address the statistical analysis within the EM framework described in *The incomplete data likelihood* section.

Equation 1 can be generalized to the case of N pedigrees,

$$
\begin{aligned}
L(\mathbf{y}^{\mathrm{F}}, \mathbf{y}^{\mathrm{N}}, \mathcal{D}^{\mathrm{F}}, \mathcal{D}^{\mathrm{N}}, \mathcal{R}|\mathbf{\Omega}) &= \prod_{i=1}^{N} L_i \propto p_{AQB}^{n_{AQB}} p_{AQb}^{n_{AQb}} \cdots p_{aqb}^{n_{aqb}} \\
&\quad \times \theta_{AQ}^{n_{\theta_{AQ}}} (1-\theta_{AQ})^{n_{\bar{\theta}_{AQ}}} \\
&\quad \times \theta_{BQ}^{n_{\theta_{BQ}}} (1-\theta_{BQ})^{n_{\bar{\theta}_{BQ}}} \\
&\quad \times \prod_{i=1}^{N} \left[\begin{array}{l} \prod_{j=1}^{N_i^{\mathrm{F}}} f(y_{ij}^{\mathrm{F}}|\mathcal{D}_{ij}^{\mathrm{F}}) \\ \quad \times \prod_{j=1}^{N_i^{\mathrm{N}}} f(y_{ij}^{\mathrm{N}}|\mathcal{D}_{ij}^{\mathrm{N}}) \end{array} \right], \qquad (1')
\end{aligned}
$$

where the likelihood of pedigree i,

$$
\begin{aligned}
L_i &= \prod_{j=1}^{N_i^{\mathrm{F}}} \left[\Pr(\mathcal{D}_{ij}^{\mathrm{F}}) f(y_{ij}^{\mathrm{F}}|\mathcal{D}_{ij}^{\mathrm{F}}) \right] \prod_{j=1}^{N_i^{\mathrm{N}}} \left[\Pr(\mathcal{D}_{ij}^{\mathrm{N}}, \mathcal{R}_{ij}|\mathcal{D}_{ij}^{\mathrm{m}}, \mathcal{D}_{ij}^{\mathrm{f}}) f(y_{ij}^{\mathrm{N}}|\mathcal{D}_{ij}^{\mathrm{N}}) \right] \\
&= \prod_{j=1}^{N_i^{\mathrm{F}}+N_i^{\mathrm{N}}} \left[\Pr(\mathcal{D}_{ij}\langle, \mathcal{R}_{ij}|\cdot\rangle) f(y_{ij}|\mathcal{D}_{ij}) \right],
\end{aligned}
$$

assuming that the rightmost is ordered as Elston and Stewart's (1971) recursive form in which $\Pr(\mathcal{D}_{ij}, \langle \mathcal{R}_{ij}|\cdot\rangle)$ represents the probability of either child j given the parental diplotypes or founder j within pedigree i; $\mathcal{D}_{ij}^{\mathrm{m}}$ and $\mathcal{D}_{ij}^{\mathrm{f}}$ are the parental diplotypes of nonfounder j within pedigree i, respectively; $\mathbf{y}^{\mathbf{F}}$ and $\mathbf{y}^{\mathbf{N}}$ are the founder and nonfounder phenotypic vectors; $\mathcal{D}^{\mathrm{F}}$ and $\mathcal{D}^{\mathrm{N}}$ are the founder and nonfounder diplotype vectors; $\mathcal{R}$ is the recombination configuration for nonfounders, respectively; N_i^{F} and N_i^{N} are the numbers of founder(s) and nonfounder(s) within pedigree i; $n_{AQB}, n_{AQb}, \ldots, n_{aqb}$ are the numbers of haplotypes AQB, AQb, $\ldots$, aqb appearing in founder diplotypes, respectively; and $n_{\theta_{AQ}}$, $n_{\bar{\theta}_{AQ}}$, $n_{\theta_{BQ}}$, and $n_{\bar{\theta}_{BQ}}$ are the numbers of recombinants and nonrecombinants between loci $\mathcal{A}$ and $\mathcal{Q}$ and between $\mathcal{B}$ and $\mathcal{Q}$ across all N pedigrees, respectively.

The maximum-likelihood estimator can be derived through differentiating the log-likelihood with respect to $\mathbf{\Omega}$ and then setting each derivative equal to 0 and solving the set of simultaneous equations. Define the identity indicators

$$
I(QQ|\mathcal{D}) = \begin{cases} 1 & \text{if } \mathcal{G}_{\mathbf{q}}(\mathcal{D}) = QQ, \\ & \quad i.e., \text{ diplotype } \mathcal{D} \text{ is compatible with} \\ & \quad \text{the genotype } QQ \\ 0 & \text{otherwise,} \end{cases}
$$

$$
I(Qq|\mathcal{D}) = \begin{cases} 1 & \text{if } \mathcal{G}_{\mathbf{q}}(\mathcal{D}) = Qq \\ 0 & \text{otherwise,} \end{cases}
$$

and

$$
I(qq|\mathcal{D}) = \begin{cases} 1 & \text{if } \mathcal{G}_{\mathbf{q}}(\mathcal{D}) = qq \\ 0 & \text{otherwise.} \end{cases}
$$

The MLEs for the likelihood (1) are

$$
\begin{aligned}
&\hat{p}_{AQB} = \frac{n_{AQB}}{4N}, \quad \hat{p}_{AQb} = \frac{n_{AQb}}{4N}, \ \ldots, \ \hat{p}_{aqb} = \frac{n_{aqb}}{4N}, \\
&\hat{\theta}_{AQ} = \frac{n_{\theta_{AQ}}}{n_{\theta_{AQ}} + n_{\bar{\theta}_{AQ}}}, \quad \hat{\theta}_{BQ} = \frac{n_{\theta_{BQ}}}{n_{\theta_{BQ}} + n_{\bar{\theta}_{BQ}}}, \\
&\hat{\mu}_G = \frac{\sum_{i=1}^{N} \left[I(G|\mathcal{D}_i^{\mathrm{m}}) y_i^{\mathrm{m}} + I(G|\mathcal{D}_i^{\mathrm{f}}) y_i^{\mathrm{f}} + \sum_{j=1}^{N_i} I(G|\mathcal{D}_{ij}^{\mathrm{o}}) y_{ij}^{\mathrm{o}} \right]}{\sum_{i=1}^{N} \left[I(G|\mathcal{D}_i^{\mathrm{m}}) + I(G|\mathcal{D}_i^{\mathrm{f}}) + \sum_{j=1}^{N_i} I(G|\mathcal{D}_{ij}^{\mathrm{o}}) \right]}, \qquad (2) \\
&\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \left[(y_i^{\mathrm{m}} - \hat{\mu}_{\mathcal{G}_{\mathbf{q}}(\mathcal{D}_i^{\mathrm{m}})})^2 + (y_i^{\mathrm{f}} - \hat{\mu}_{\mathcal{G}_{\mathbf{q}}(\mathcal{D}_i^{\mathrm{f}})})^2 + \sum_{j=1}^{N_i} (y_{ij}^{\mathrm{o}} - \hat{\mu}_{\mathcal{G}_{\mathbf{q}}(\mathcal{D}_{ij}^{\mathrm{o}})})^2 \right]}{2N + \sum_{i=1}^{N} N_i},
\end{aligned}
$$

for quantitative traits, and

$$\hat{f}(y|G) = \left(\sum_{i=1}^{N}\left[I(G|\mathcal{D}_i^{\mathrm{m}})I(y=y_i^{\mathrm{m}}) + I(G|\mathcal{D}_i^{\mathrm{f}})I(y=y_i^{\mathrm{f}}) + \sum_{j=1}^{N_i} I(G|\mathcal{D}_{ij}^{\mathrm{o}})I(y=y_{ij}^{\mathrm{o}})\right]\right) \Big/ \left(\sum_{i=1}^{N}\left[I(G|\mathcal{D}_i^{\mathrm{m}}) + I(G|\mathcal{D}_i^{\mathrm{f}}) + \sum_{j=1}^{N_i} I(G|\mathcal{D}_{ij}^{\mathrm{o}})\right]\right),$$

for categorical traits, where $G \in \{QQ, Qq, qq\}$; and indicators $I(y=y_i^{\mathrm{m}})$, $I(y=y_i^{\mathrm{f}})$, and $I(y=y_{ij}^{\mathrm{o}})$ are 1 when $y=y_i^{\mathrm{m}}$, $y=y_i^{\mathrm{f}}$, and $y=y_{ij}^{\mathrm{o}}$, respectively, and 0 otherwise. The MLEs of the recombination parameters for likelihood (1′) are the same as those for likelihood (1), and the other MLEs have similar forms,

$$\hat{p}_{AQB} = \frac{n_{AQB}}{2\sum_{i=1}^{N} N_i^{\mathrm{F}}}, \quad \hat{p}_{AQb} = \frac{n_{AQb}}{2\sum_{i=1}^{N} N_i^{\mathrm{F}}}, \ \cdots, \quad \hat{p}_{aqb} = \frac{n_{aqb}}{2\sum_{i=1}^{N} N_i^{\mathrm{F}}},$$

$$\hat{\mu}_G = \frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{N_i^{\mathrm{F}}} I(G|\mathcal{D}_{ij}^{\mathrm{F}})y_{ij}^{\mathrm{F}} + \sum_{j=1}^{N_i^{\mathrm{N}}} I(G|\mathcal{D}_{ij}^{\mathrm{N}})y_{ij}^{\mathrm{N}}\right]}{\sum_{i=1}^{N}\left[\sum_{j=1}^{N_i^{\mathrm{F}}} I(G|\mathcal{D}_{ij}^{\mathrm{F}}) + \sum_{j=1}^{N_i^{\mathrm{N}}} I(G|\mathcal{D}_{ij}^{\mathrm{N}})\right]}, \tag{2′}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{N_i^{\mathrm{F}}}(y_{ij}^{\mathrm{F}} - \hat{\mu}_{\mathcal{G}_{\mathbf{q}}(\mathcal{D}_{ij}^{\mathrm{F}})})^2 + \sum_{j=1}^{N_i^{\mathrm{N}}}(y_{ij}^{\mathrm{N}} - \hat{\mu}_{\mathcal{G}_{\mathbf{q}}(\mathcal{D}_{ij}^{\mathrm{N}})})^2\right]}{\sum_{i=1}^{N}(N_i^{\mathrm{F}} + N_i^{\mathrm{N}})},$$

for quantitative traits, and

$$\hat{f}(y|G) = \frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{N_i^{\mathrm{F}}} I(G|\mathcal{D}_{ij}^{\mathrm{F}})I(y=y_{ij}^{\mathrm{F}}) + \sum_{j=1}^{N_i^{\mathrm{N}}} I(G|\mathcal{D}_{ij}^{\mathrm{N}})I(y=y_{ij}^{\mathrm{N}})\right]}{\sum_{i=1}^{N}\left[\sum_{j=1}^{N_i^{\mathrm{F}}} I(G|\mathcal{D}_{ij}^{\mathrm{F}}) + \sum_{j=1}^{N_i^{\mathrm{N}}} I(G|\mathcal{D}_{ij}^{\mathrm{N}})\right]},$$

for category traits.

Unlike the traditional approach, for flexibility we make here no assumption such as that the recombination fraction between the two markers can be known *a priori*. If the recombination fraction between the two markers (θ_{AB}) is available, however, the corresponding terms with respect to one of the recombination fractions, θ_{AQ} and θ_{BQ}, will disappear from the above estimation procedure since any one of the two is a function of the other one and of θ_{AB}. A grid search procedure can also be used for estimating QTL position on the basis of the preceding methodology.

The incomplete data likelihood: In practice, only marker genotype and phenotype data are observed, whereas the data on diplotypes, recombination events, and QTL genotypes are hidden. The observed data are mixtures of component complete data, and the statistical analysis becomes a typical mixture issue. Let us go back to the above example again and assume that only marker genotypes *AABB* (M_i^{m}), *AABb* (M_i^{f}), *AABB* (M_{i1}^{o}), and *AABb* (M_{i2}^{o}) and phenotypes y_i^{m}, y_i^{f}, y_{i1}^{o}, and y_{i2}^{o} are available for the mother, father, and two children of family *i*, respectively. Now M_i^{m} is a mixture of diplotypes *AQB/AQB*, *AQB/AqB*, and *AqB/AqB*; and M_i^{f} is composed of diplotypes *AQB/AQb*, *AQB/Aqb*, *AQb/AqB*, and *AqB/Aqb*; and both M_{i1}^{o} and M_{i2}^{o} also consist of unidentified diplotype(s) together with recombination configuration(s) nested within the paired parental diplotypes. The likelihood can be formulated as

$$\begin{aligned}
L_i &= L(y_i^{\mathrm{m}}, y_i^{\mathrm{f}}, y_{i1}^{\mathrm{o}}, y_{i2}^{\mathrm{o}}, M_i^{\mathrm{m}}, M_i^{\mathrm{f}}, M_{i1}^{\mathrm{o}}, M_{i2}^{\mathrm{o}}|\boldsymbol{\Omega}) \\
&= \Pr(AQB/AQB)\Pr(AQB/AQb)f(y_i^{\mathrm{m}}|QQ)f(y_i^{\mathrm{f}}|QQ) \\
&\quad \times \sum_{*\in\{N,R\}} \Pr(AQB/AQB, **/**|AQB/AQB, AQB/AQb)f(y_{i1}^{\mathrm{o}}|QQ) \\
&\quad \times \sum_{*\in\{N,R\}} \Pr(AQB/AQb, **/**|AQB/AQB, AQB/AQb)f(y_{i2}^{\mathrm{o}}|QQ) \\
&\quad + \Pr(AQB/AQB)\Pr(AQB/Aqb)f(y_i^{\mathrm{m}}|QQ)f(y_i^{\mathrm{f}}|Qq) \\
&\quad \times \left[\sum_{*\in\{N,R\}} \Pr(AQB/AQB, **/*N|AQB/AQB, AQB/Aqb)f(y_{i1}^{\mathrm{o}}|QQ) + \sum_{*\in\{N,R\}} \Pr(AQB/AqB, **/*R|AQB/AQB, AQB/Aqb)f(y_{i1}^{\mathrm{o}}|Qq)\right] \\
&\quad \times \left[\sum_{*\in\{N,R\}} \Pr(AQB/AQb, **/*R|AQB/AQB, AQB/Aqb)f(y_{i2}^{\mathrm{o}}|QQ) + \sum_{*\in\{N,R\}} \Pr(AQB/Aqb, **/*N|AQB/AQB, AQB/Aqb)f(y_{i2}^{\mathrm{o}}|Qq)\right] \\
&\quad + \cdots \\
&= L_i(AQB/AQB, AQB/AQb) + L_i(AQB/AQB, AQB/Aqb) + \cdots \\
&= \sum_{\mathcal{D}_i^{\mathrm{m}},\mathcal{D}_i^{\mathrm{f}}} L_i(\mathcal{D}_i^{\mathrm{m}}, \mathcal{D}_i^{\mathrm{f}}),
\end{aligned}$$

where $\sum_{*\in\{N,R\}}$ denotes summation over all recombination configuration(s) by taking "*" as either recombination or nonrecombination that is compatible with parent and child diplotypes; $L_i(AQB/AQB, AQB/AQb)$, $L_i(AQB/AQB, AQB/Aqb)$, ..., are probabilities of the mother and father of family *i* with diplotypes *AQB/AQB* and *AQB/AQb*, *AQB/AQB* and *AQB/Aqb*, ..., respectively; and $\sum_{(\mathcal{D}_i^{\mathrm{m}},\mathcal{D}_i^{\mathrm{f}})}$ denotes summation over all pairs of $(\mathcal{D}_i^{\mathrm{m}}, \mathcal{D}_i^{\mathrm{f}})$ compatible with the observed marker phenotypes in family *i*. The partial derivative of the log-likelihood of family *i* is

$$\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{\Omega}}\ln L(y_i^{\mathrm{m}}, y_i^{\mathrm{f}}, y_{i1}^{\mathrm{o}}, y_{i2}^{\mathrm{o}}, M_i^{\mathrm{m}}, M_i^{\mathrm{f}}, M_{i1}^{\mathrm{o}}, M_{i2}^{\mathrm{o}}|\boldsymbol{\Omega}) \\
&= \boldsymbol{\pi}^{i}_{(AQB/AQB,AQB/AQb)}\left[\begin{array}{l}\frac{\partial}{\partial\boldsymbol{\Omega}_{\mathbf{P}}}\ln\Pr(AQB/AQB) \\ \quad + \frac{\partial}{\partial\boldsymbol{\Omega}_{\mathbf{P}}}\ln\Pr(AQB/AQb) \\ \quad + \frac{\partial}{\partial\boldsymbol{\Omega}_{\mathbf{Q}}}\ln f(y_i^{\mathrm{m}}|QQ) \\ \quad + \frac{\partial}{\partial\boldsymbol{\Omega}_{\mathbf{Q}}}\ln f(y_i^{\mathrm{f}}|QQ)\end{array}\right] \\
&\quad + \sum_{*\in\{N,R\}} \boldsymbol{\pi}^{i1}_{(AQB/AQB,**/**|AQB/AQB,AQB/AQb)} \\
&\quad \times \left[\begin{array}{l}\frac{\partial}{\partial\boldsymbol{\Omega}_{\mathbf{R}}}\ln\Pr(AQB/AQB, **/**|AQB/AQB, AQB/AQb) \\ \quad + \frac{\partial}{\partial\boldsymbol{\Omega}_{\mathbf{Q}}}\ln f(y_{i1}^{\mathrm{o}}|QQ)\end{array}\right] \\
&\quad + \sum_{*\in\{N,R\}} \boldsymbol{\pi}^{i2}_{(AQB/AQb,**/**|AQB/AQB,AQB/AQb)} \\
&\quad \times \left[\begin{array}{l}\frac{\partial}{\partial\boldsymbol{\Omega}_{R}}\ln\Pr(AQB/AQb, **/**|AQB/AQB, AQB/AQb) \\ \quad + \frac{\partial}{\partial\boldsymbol{\Omega}_{Q}}\ln f(y_{i2}^{\mathrm{o}}|QQ)\end{array}\right] \\
&\quad + \cdots,
\end{aligned}$$

where $\boldsymbol{\pi}^{i}_{(\mathcal{D}_i^{\mathrm{m}},\mathcal{D}_i^{\mathrm{f}})}$ and $\boldsymbol{\pi}^{ij}_{(\mathcal{D}_{ij}^{\mathrm{o}},\mathcal{R}_{ij}|\mathcal{D}_i^{\mathrm{m}},\mathcal{D}_i^{\mathrm{f}})}$ are the posterior probabilities that the mother and father of family *i* have diplotypes $\mathcal{D}_i^{\mathrm{m}}$ and $\mathcal{D}_i^{\mathrm{f}}$ and that child *j* from family *i* has diplotype $\mathcal{D}_{ij}^{\mathrm{o}}$ and reduced recombination $\mathcal{R}_{ij}$ produced by the mother and father diplotypes $\mathcal{D}_i^{\mathrm{m}}$ and $\mathcal{D}_i^{\mathrm{f}}$, respectively; *e.g.*,

$$\boldsymbol{\pi}^{i}_{(AQB/AQB,AQB/Aqb)} = \frac{L_i(AQB/AQB, AQB/Aqb)}{L_i}$$

and

$$\begin{aligned}\pi^{i1}_{(AQB/AQB,NN/NN|AQB/AQB,AQB/Aqb)}&=\pi^{i}_{(AQB/AQB,AQB/Aqb)}\\&\times[\Pr(AQB/AQB,\ NN/NN|AQB/AQB,\ AQB/Aqb)\\&\times f(y^{\mathrm{o}}_{i1}|QQ)]/\Bigg[\sum_{*\in\{N,R\}}\Pr(AQB/AQB,\ **/*N|AQB/AQB,\ AQB/Aqb)\\&\times f(y^{\mathrm{o}}_{i1}|QQ)\\&+\sum_{*\in\{N,R\}}\Pr(AQB/AqB,\ **/*R|AQB/AQB,\ AQB/Aqb)\\&\times f(y^{\mathrm{o}}_{i1}|Qq)\Big].\end{aligned}$$

The grand likelihood of the incomplete data, including the phenotype ($\mathbf{y}$) and marker information ($\mathbf{M}$), can be represented as

$$\begin{aligned}&L(\mathbf{y}^{\mathrm{m}},\mathbf{y}^{\mathrm{f}},\mathbf{y}^{\mathrm{o}},\mathbf{M}^{\mathrm{m}},\mathbf{M}^{\mathrm{f}},\mathbf{M}^{\mathrm{o}}|\boldsymbol{\Omega})\\&\quad\propto\prod_{i=1}^{N}L(y^{\mathrm{m}}_i,y^{\mathrm{f}}_i,y^{\mathrm{o}}_{i1},y^{\mathrm{o}}_{i2},\ \dots\ ,M^{\mathrm{m}}_i,M^{\mathrm{f}}_i,M^{\mathrm{o}}_{i1},M^{\mathrm{o}}_{i2},|\boldsymbol{\Omega}),\end{aligned}\tag{3}$$

where $\mathbf{M}^{\mathrm{m}}$, $\mathbf{M}^{\mathrm{f}}$, and $\mathbf{M}^{\mathrm{o}}$ are the marker genotypes of the mothers, fathers, and children, respectively.

Differentiating the log-likelihood of Equation 3 leads to

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\Omega}}\ln L(\mathbf{y}^{\mathrm{m}},\mathbf{y}^{\mathrm{f}},\mathbf{y}^{\mathrm{o}},\mathbf{M}^{\mathrm{m}},\mathbf{M}^{\mathrm{f}},\mathbf{M}^{\mathrm{o}}|\boldsymbol{\Omega})\\&\quad=\sum_{i=1}^{N}\frac{\partial\ln L(y^{\mathrm{m}}_i,y^{\mathrm{f}}_i,y^{\mathrm{o}}_{i1},y^{\mathrm{o}}_{i2},\ \dots\ ,M^{\mathrm{m}}_i,M^{\mathrm{f}}_i,M^{\mathrm{o}}_{i1},M^{\mathrm{o}}_{i2},\dots|\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}}\\&\quad=n^{*}_{AQB}\frac{\partial\ln p_{AQB}}{\partial\boldsymbol{\Omega}_{\mathbf{P}}}+n^{*}_{AQb}\frac{\partial\ln p_{AQb}}{\partial\boldsymbol{\Omega}_{\mathbf{P}}}+\cdots n^{*}_{aqb}\frac{\partial\ln p_{aqb}}{\partial\boldsymbol{\Omega}_{\mathbf{P}}}\\&\qquad+n^{*}_{\theta_{AQ}}\frac{\partial\ln\theta_{AQ}}{\partial\boldsymbol{\Omega}_{\mathbf{R}}}+n^{*}_{\bar\theta_{AQ}}\frac{\partial\ln(1-\theta_{AQ})}{\partial\boldsymbol{\Omega}_{\mathbf{R}}}+n^{*}_{\theta_{BQ}}\frac{\partial\ln\theta_{BQ}}{\partial\boldsymbol{\Omega}_{\mathbf{R}}}\\&\qquad+n^{*}_{\bar\theta_{BQ}}\frac{\partial\ln(1-\theta_{BQ})}{\partial\boldsymbol{\Omega}_{\mathbf{R}}}\\&\qquad+\sum_{i=1}^{N}\sum_{\mathcal{D}_\xi,\mathcal{D}_\varepsilon}\left\{\begin{array}{l}\pi^{i}_{(\mathcal{D}_\xi,\mathcal{D}_\varepsilon)}\left[\dfrac{\partial\ln f(y^{\mathrm{m}}_i|\mathcal{D}_\xi)}{\partial\boldsymbol{\Omega}_{\mathbf{Q}}}+\dfrac{\partial\ln f(y^{\mathrm{f}}_i|\mathcal{D}_\varepsilon)}{\partial\boldsymbol{\Omega}_{\mathbf{Q}}}\right]\\\quad+\sum_{j=1}^{N_i}\sum_{\mathcal{D}_\xi,\mathcal{R}_\tau}\pi^{ij}_{(\mathcal{D}_\xi,\mathcal{R}_\tau|\mathcal{D}_\xi,\mathcal{D}_\varepsilon)}\dfrac{\partial\ln f(y^{\mathrm{o}}_{ij}|\mathcal{D}_\zeta)}{\partial\boldsymbol{\Omega}_{\mathbf{Q}}}\end{array}\right\},\end{aligned}\tag{4}$$

where n^{*}_{AQB}, n^{*}_{AQb}, $\dots$, n^{*}_{aqb} are the expected numbers of haplotypes AQB, AQb, $\dots$, and aqb, respectively; $n^{*}_{\theta_{AQ}}$, $n^{*}_{\theta_{BQ}}$, $n^{*}_{\bar\theta_{AQ}}$, and $n^{*}_{\bar\theta_{BQ}}$ are the expected numbers of recombinants and nonrecombinants between $\mathcal{A}$ and $\mathcal{Q}$ and between $\mathcal{B}$ and $\mathcal{Q}$, respectively; and sums are taken over all diplotypes and recombination configurations consistent with the marker genotypes.

Similarly, the pedigree-based likelihood is

$$\begin{aligned}L(\mathbf{y}^{\mathrm{F}},\mathbf{y}^{\mathrm{N}},\mathbf{M}^{\mathrm{F}},\mathbf{M}^{\mathrm{N}}|\boldsymbol{\Omega})&\propto\prod_{i=1}^{N}L(y^{\mathrm{F}}_i,y^{\mathrm{N}}_i,M^{\mathrm{F}}_i,M^{\mathrm{N}}_i|\boldsymbol{\Omega})\\&=\prod_{i=1}^{N}\prod_{j=1}^{N^{\mathrm{F}}_i}\sum_{\mathcal{D}_\xi}[\Pr(\mathcal{D}_\xi|M_{ij})f(y_{ij}|\mathcal{D}_\xi)]\\&\quad\times\prod_{j=1}^{N^{\mathrm{N}}_i}\sum_{\mathcal{D}_\varepsilon,\mathcal{R}_\tau}[\Pr(\mathcal{D}_\varepsilon,\mathcal{R}_\tau|\mathcal{D}^{\mathrm{m}}_{ij},\mathcal{D}^{\mathrm{f}}_{ij})f(y^{\mathrm{N}}_{ij}|\mathcal{D}_\varepsilon)]\\&=\prod_{i=1}^{N}\prod_{j=1}^{N^{\mathrm{F}}_i+N^{\mathrm{N}}_i}\sum_{\mathcal{D}_\xi<,\mathcal{R}_\tau>}[\Pr(\mathcal{D}_\xi<,\mathcal{R}_\tau|\cdot>)f(y_{ij}|\mathcal{D}_\xi)],\end{aligned}\tag{3′}$$

where $\mathbf{M}^{\mathrm{F}}$ and $\mathbf{M}^{\mathrm{N}}$ are the marker genotypes of founder(s) and nonfounder(s), respectively, the last line is placed in a recursive order, and $\sum$ denotes summation over all diplotype(s) and/or recombination configuration(s) compatible with the observed data. The partial derivative is

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\Omega}}\ln L(\mathbf{y}^{\mathrm{F}},\mathbf{y}^{\mathrm{N}},\mathbf{M}^{\mathrm{F}},\mathbf{M}^{\mathrm{N}}|\boldsymbol{\Omega})&=n^{*}_{AQB}\frac{\partial\ln p_{AQB}}{\partial\boldsymbol{\Omega}_{\mathbf{P}}}+n^{*}_{AQb}\frac{\partial\ln p_{AQb}}{\partial\boldsymbol{\Omega}_{\mathbf{P}}}\\&\quad+\cdots+n^{*}_{aqb}\frac{\partial\ln p_{aqb}}{\partial\boldsymbol{\Omega}_{\mathbf{P}}}\\&\quad+n^{*}_{\theta_{AQ}}\frac{\partial\ln\theta_{AQ}}{\partial\boldsymbol{\Omega}_{\mathbf{R}}}+n^{*}_{\bar\theta_{AQ}}\frac{\partial\ln(1-\theta_{AQ})}{\partial\boldsymbol{\Omega}_{\mathbf{R}}}\\&\quad+n^{*}_{\theta_{BQ}}\frac{\partial\ln\theta_{BQ}}{\partial\boldsymbol{\Omega}_{\mathbf{R}}}+n^{*}_{\bar\theta_{BQ}}\frac{\partial\ln(1-\theta_{BQ})}{\partial\boldsymbol{\Omega}_{\mathbf{R}}}\\&\quad+\sum_{i=1}^{N}\left[\begin{array}{l}\sum_{j=1}^{N^{\mathrm{F}}_i}\sum_{\mathcal{D}_\xi}\pi^{ij}_{(\mathcal{D}_\xi)}\frac{\partial\ln f(y^{\mathrm{F}}_{ij}|\mathcal{D}_\xi)}{\partial\boldsymbol{\Omega}_{\mathbf{Q}}}\\\quad+\sum_{j=1}^{N^{\mathrm{N}}_i}\sum_{\mathcal{D}_\xi,\mathcal{D}_\varepsilon}\sum_{\mathcal{D}_\zeta,\mathcal{R}_\tau}\pi^{ij}_{(\mathcal{D}_\zeta,\mathcal{R}_\tau|\mathcal{D}_\xi,\mathcal{D}_\varepsilon)}\frac{\partial\ln f(y^{\mathrm{N}}_{ij}|\mathcal{D}_\zeta)}{\partial\boldsymbol{\Omega}_{\mathbf{Q}}}\end{array}\right].\end{aligned}\tag{4′}$$

We can adopt the *peeling algorithm* (Elston and Stewart 1971) to calculate the likelihood and the posterior probabilities. Under the assumption of linkage equilibrium, our approach reduces to an EM version of Elston and Stewart's (1971) algorithm. For pedigree(s) with loop(s), we can use Lange and Elston's (1975) method to break the loop(s).

We implement the EM algorithm (Dempster *et al.* 1977) to estimate the parameters of the likelihood function, *i.e.*, haplotype frequencies $\boldsymbol{\Omega}_P$, QTL genotypic effects and residual variance or penetrances $\boldsymbol{\Omega}_Q$, and recombination fractions $\boldsymbol{\Omega}_R$. In the E-step, we update the posterior probabilities and expected numbers conditional on the initial values or the estimates of the current iteration. In the M-step, substituting expected numbers n^{*}_{AQB}, n^{*}_{AQb}, $\dots$, n^{*}_{aqb}, $n^{*}_{\theta_{AQ}}$, $n^{*}_{\theta_{BQ}}$, $n^{*}_{\bar\theta_{AQ}}$, and $n^{*}_{\bar\theta_{BQ}}$ and posterior probabilities $\sum_{\mathcal{D}_\xi,\mathcal{D}_\zeta}I(G|\mathcal{D}_\xi)\pi^{i}_{(\mathcal{D}_\xi,\mathcal{D}_\zeta)}$, $\sum_{\mathcal{D}_\xi,\mathcal{D}_\zeta}I(G|\mathcal{D}_\zeta)\pi^{i}_{(\mathcal{D}_\xi,\mathcal{D}_\zeta)}$, and $\sum_{\mathcal{D}_\xi,\mathcal{D}_\zeta}\sum_{\mathcal{D}_\varsigma,\mathcal{R}_\tau}I(G|\mathcal{D}_\varsigma)\cdot\pi^{ij}_{(\mathcal{D}_\varsigma,\mathcal{R}_\tau|\mathcal{D}_\xi,\mathcal{D}_\zeta)}$ for $I(G|\mathcal{D}^{\mathrm{f}}_i)$, $I(G|\mathcal{D}^{\mathrm{m}}_i)$, and $I(G|\mathcal{D}^{\mathrm{o}}_{ij})$ in Equations 2 for likelihood (3), respectively, $G\in\{QQ, Qq, qq\}$, we compute the next cycle of MLEs of the unknown parameters. Likewise, we perform a similar M-step in (2′) for the pedigree-based likelihood (3′). These two steps are repeated until convergence is attained. Allele frequencies and linkage disequilibria, QTL additive and dominance effects, and relative locations on the chromosome can be calculated from the haplotype frequencies, QTL genotypic effects, and recombination fractions, respectively.

The asymptotic variance–covariance matrix of the MLEs: Louis' (1982) procedure or the supplemented EM (SEM) (Meng and Rubin 1991) that embeds the computation of the observed information within the EM iteration can be adopted to obtain the asymptotic variance–covariance matrix for MLEs of haplotype frequencies, genotypic effects, residual variance, and recombination fraction(s). In our computer program we use the improved equations of Louis' (1982) method, Lou *et al.*'s (2005) (C3) and (C5), to compute the

observed information matrix of the parameters (*i.e.*, the haplotype frequencies, penetrance or genotypic effects plus residual variance, and recombination fractions). The information on other parameters can be calculated with that for these basic parameters. The variance–covariance matrix for genetic effects and allele frequencies can be calculated easily since they are linear functions of the haplotype frequencies or genotypic effects. The approximate variances of the linkage disequilibria can be found by the *delta method*, based on their Taylor series expansions. If the parameter vector φ is a function of the basic parameter ϕ, *i.e.*, $\varphi = f(\phi)$, then the approximate variance–covariance of $\hat{\varphi} = f(\hat{\phi})$ is given by

$$\mathrm{Var}(\hat{\varphi}) = \frac{\partial f(\phi)}{\partial \phi}\mathrm{Var}(\hat{\phi})\frac{\partial f(\phi)}{\partial \phi^{\mathrm{T}}}. \qquad (5)$$

For example,

$$\mathrm{Var}\begin{pmatrix}\hat{\delta}_{AB}\\ \hat{\delta}_{AQ}\\ \hat{\delta}_{BQ}\\ \hat{\delta}_{AQB}\end{pmatrix} \triangleq \hat{\mathbf{S}}^{\mathrm{T}} \times \hat{\mathrm{Var}}\begin{pmatrix}\hat{p}_{AQB}\\ \hat{p}_{AQb}\\ \hat{p}_{AqB}\\ \hat{p}_{Aqb}\\ \hat{p}_{aQB}\\ \hat{p}_{aQb}\\ \hat{p}_{aqB}\end{pmatrix} \times \hat{\mathbf{S}},$$

where

$$\hat{\mathbf{S}} = \begin{pmatrix} 1-\hat{p}_A-\hat{p}_B & 1-\hat{p}_A-\hat{p}_Q & 1-\hat{p}_B-\hat{p}_Q \\ -\hat{p}_B & 1-\hat{p}_A-\hat{p}_Q & -\hat{p}_B \\ 1-\hat{p}_A-\hat{p}_B & -\hat{p}_Q & -\hat{p}_Q \\ -\hat{p}_B & -\hat{p}_Q & 0 \\ -\hat{p}_A & -\hat{p}_A & 1-\hat{p}_B-\hat{p}_Q \\ 0 & -\hat{p}_A & -\hat{p}_B \\ -\hat{p}_A & 0 & -\hat{p}_Q \end{pmatrix}.$$

$$\begin{pmatrix} 1-\hat{D}_{AB}-\hat{D}_{AQ}-\hat{D}_{BQ}-\hat{p}_A-\hat{p}_B-\hat{p}_Q+2\hat{p}_A\hat{p}_B+2\hat{p}_A\hat{p}_Q+2\hat{p}_B\hat{p}_Q \\ 2\hat{p}_A\hat{p}_B+2\hat{p}_B\hat{p}_Q-\hat{D}_{AB}-\hat{D}_{BQ}-\hat{p}_B \\ 2\hat{p}_A\hat{p}_Q+2\hat{p}_B\hat{p}_Q-\hat{D}_{AQ}-\hat{D}_{BQ}-\hat{p}_Q \\ 2\hat{p}_B\hat{p}_Q-\hat{D}_{BQ} \\ 2\hat{p}_A\hat{p}_B+2\hat{p}_A\hat{p}_Q-\hat{D}_{AB}-\hat{D}_{AQ}-\hat{p}_A \\ 2\hat{p}_A\hat{p}_B-\hat{D}_{AB} \\ 2\hat{p}_A\hat{p}_Q-\hat{D}_{AQ} \end{pmatrix}.$$

Hypothesis testing: The following hypotheses are tested sequentially: (1) the existence of a trait gene and (2) various submodel hypotheses. The existence of a trait gene with significant effects can be tested by calculating a log-likelihood ratio (LR) test statistic under the null (H_0: there is no trait-causing gene) and alternative hypotheses (H_1: there is a trait-causing gene) as

$$\mathrm{LR} = -2[\log L_0(\mu_{QQ} = \mu_{Qq} = \mu_{qq} = \tilde{\mu}, \tilde{\sigma}^2, \tilde{\Omega}_P, \tilde{\Omega}_R) - \log L_1(\hat{\mathbf{\Omega}})],$$

for quantitative traits and

$$\mathrm{LR} = -2[\log L_0[f(y|QQ) = f(y|Qq) = f(y|qq) = \tilde{f}(y|\cdot), \tilde{\Omega}_P, \tilde{\Omega}_R] - \log L_1(\hat{\mathbf{\Omega}})]$$

for categorical traits. The LR under the null hypothesis is asymptotically χ^2-distributed with corresponding degrees of freedom for a fixed set of frequencies and relative position of the putative gene. However, because these are nuisance parameters under H_0, the regularity conditions required for the χ^2-distribution of the LR statistic are violated. Parametric or nonparametric bootstrap (*e.g.*, the permutation procedure proposed by Churchill and Doerge, 1994) can be adopted to determine a critical threshold for declaring the presence of a gene at a given significance level.

After rejecting the hypothesis of no gene, the tests for particular subsets of hypotheses regarding gene action mode, gene position, and/or LD coefficient(s) can be conducted in tandem with the corresponding LR statistics that are approximately distributed as χ^2-statistics with degrees of freedom equal to the relevant numbers of parameters being tested.

Benefiting from making full use of both complementary components of information on correlated transmission within pedigrees and correlated occurrence at the population level, the proposed approach is expected to have greater analytical accuracy and testing power. To validate our theoretical expectation, we conducted a series of simulations under a variety of disease models and degrees of LD to compare the performance of three methods: an FBAT, pure linkage (PL) analysis, and the combined linkage and association analysis (LLD).

SIMULATION STUDIES

A model with two diallelic loci, one marker and one disease gene each with a minor allele frequency of 0.4, was considered in our simulation studies. LLD was run by a computer program written in the C++ language, while PL analysis was performed by the EM version of Elston and Stewart's (1971) algorithm. Average MLE, mean square error (MSE), and the power of both LLD and PL were computed on the basis of 200 simulations for each case. Power calculation of the FBAT was implemented with the PBAT software package on the basis of simulation using the default choice (Lange and Laird 2002; Lange *et al.* 2002). Unless otherwise stated, all powers were evaluated at the 0.05 significance level for a null hypothesis of no linkage. The complete details of the scenarios used in the simulations are given in the relevant text and tables of this section.

To confirm that PL analysis may result in a biased estimation in the presence of association while our new approach can remedy this limitation, we first conducted a set of simulations for a comparison between LLD and PL. Although such a case may represent an extreme one, for full exposition we concentrate here on a completely penetrant codominant disease model and, theoretically, the general conclusions from this will also be valid for complex models. TDT-type (including parent and child marker genotypes and child phenotypes) and sib-type (including sibling marker genotypes and phenotypes) data were simulated on a sample consisting of 300 nuclear families with two children and 200 families

TABLE 2

Average MLEs (and root MSEs) of the recombination fraction (θ) from pure linkage analysis (PL) and combined linkage and association analysis (LLD) at two levels of LD for TDT-type and sib-type designs, respectively

Design	δ (δ′)	θ	LLD	PL
TDT-type	0.1 (0.417)	0.05	0.049 (0.014)	0.035 (0.020)
		0.2	0.197 (0.033)	0.178 (0.049)
	0.2 (0.833)	0.05	0.050 (0.009)	0.020 (0.030)
		0.2	0.199 (0.017)	0.135 (0.069)
Sib-type	0.1 (0.417)	0.05	0.052 (0.023)	0.017 (0.036)
		0.2	0.203 (0.043)	0.142 (0.072)
	0.2 (0.833)	0.05	0.051 (0.013)	0.005 (0.046)
		0.2	0.201 (0.023)	0.051 (0.150)

with three children at two LD levels, $\delta = 0.1$ (normalized LD, $\delta' = 0.417$) and $\delta = 0.2$ ($\delta' = 0.833$), and two linkage levels, $\theta = 0.05$ and $\theta = 0.2$, respectively. Only the results on the MLE and MSE of the recombination fraction are shown in Table 2, since the MLEs of the other parameters, such as allele frequencies and LD coefficient (for LLD), have an excellent accuracy and the statistical power is very high. Table 2 shows that PL yields a large bias ($\hat{\theta} - \theta$) in both TDT-type and sib-type designs. For example, the bias and the root MSE of the estimated recombination fraction are 0.065 and 0.069 for TDT-type design and 0.149 and 0.150 for sib-type design, respectively, when true parameters are $\theta = 0.2$ and $\delta = 0.2$. This implies that the result from linkage-only analysis is less reliable when association is present. As expected, however, LLD has highly precise estimation. All the absolute values of the bias from LLD are <5% of the parameter values, a conventional criterion for unbiased estimation, and all the MSEs are much less than their counterparts from PL. The bias and the root MSE are −0.001 and 0.017 for the TDT-type design and 0.001 and 0.023 for the sib-type design, respectively, when $\theta = 0.2$ and $\delta = 0.2$.

To demonstrate that LLD can give an unbiased estimate of the recombination fraction and further test an arbitrary null hypothesis, say H_0: $\theta = 0.1$, in such a way that it has an advantage over FBATs in being capable of identifying tight linkage, we carried out simulations on the basis of a classic TDT-type design consisting of 500 nuclear families with a single child per family under a fully penetrant codominant model. As before, we considered two LD levels, $\delta = 0.1$ ($\delta' = 0.417$) and $\delta = 0.2$ ($\delta' = 0.833$), and two tight linkage levels, $\theta = 0$ and $\theta = 0.05$, respectively. Powers were calculated for the hypotheses H_0: $\theta = 0.5$ and H_0: $\theta \geq 0.1$, respectively, in LLD analysis. The MLE and MSE of the recombination fraction and the corresponding powers are presented in Table 3. As shown in Table 3, LLD gives an accurate estimate and high power for both null hypotheses at $\delta' = 0.833$; *e.g.*, the bias and the root MSE are 0.007 and 0.014, the powers for both H_0: $\theta = 0.5$ and H_0: $\theta \geq 0.1$ are 1.0, in the case of $\theta = 0$, and the bias and the root MSE are −0.004 and 0.024, and the powers are 0.995 and 0.645, in the case of $\theta = 0.05$, respectively. LLD has reasonable estimation accuracy and test power at $\delta' = 0.417$. These results suggest that LLD can offer the possibility of distinguishing strong association and loose linkage from weak association and tight linkage, even in the case of only one child per family.

TABLE 3

Average MLEs (and root MSEs) of the recombination fraction (θ) and powers for two null hypotheses (H_0: $\theta = 0.5$ and H_0: $\theta \geq 0.1$) from combined linkage and association analysis (LLD) at two levels of LD for the TDT design

δ (δ′)	θ	LLD	H_0: $\theta = 0.5$	H_0: $\theta \geq 0.1$
0.1 (0.417)	0	0.034 (0.059)	1.000	0.320
	0.05	0.062 (0.063)	1.000	0.110
0.2 (0.833)	0	0.007 (0.014)	1.000	0.995
	0.05	0.046 (0.024)	1.000	0.645

Next we consider a more common case where the disease gene affects a quantitative phenotype. TDT-type data were generated on a sample that consists of 300 families with two children each and 200 families with three children each under an additive model (no dominance effect, *i.e.*, $\mu_{QQ} = \mu + a$, $\mu_{Qq} = \mu$, and $\mu_{qq} = \mu - a$, where μ and a are the mean and additive effect, respectively). We assumed that a marker locus is completely linked to the disease susceptibility locus but with varying degrees of LD (from 0 to 0.1) and heritability (from 0.1 to 0.4). The results of the power comparison of the three methods are summarized in Figures 1 and 2, while only the estimated parameters from LLD and PL are shown in Table 4, because the nonparametric FBAT approach cannot perform parameter estimation. Figure 1 shows power plotted against LD, where a, b, c, and d are for heritabilities 0.1, 0.2, 0.3 and 0.4, respectively. Figure 2 shows power plotted against heritability, where a, b, c, d, and e are for no LD ($\delta = 0$), $\delta = 0.025$ ($\delta' = 0.104$), $\delta = 0.05$ ($\delta' = 0.208$), $\delta = 0.075$ ($\delta' = 0.313$), and $\delta = 0.1$ ($\delta' = 0.418$), respectively.

Clearly, the power profiles shown in Figures 1 and 2 support our expectation. As the degree of LD increases, so does the power of the FBAT and LLD, whereas that of PL is almost unchanged or increases little (Figure 1, a–d). The power also increases with heritability for most cases, but when there is no LD, the FBAT has no power regardless of the value of the heritability (Figure 2a). Generally speaking, it appears that LLD is the most powerful, followed by PL and then the FBAT when LD is absent or weak ($\delta' < \sim 0.2$; Figure 2, a and b) or by the FBAT and then PL when LD is strong (Figure 2, c–e). Other than the cases of no LD, where PL has power close to that of LLD, LLD is much more powerful than PL. Also, LLD always performs better than the FBAT, even under situations with strong LD ($\delta' \geq 0.313$), where the

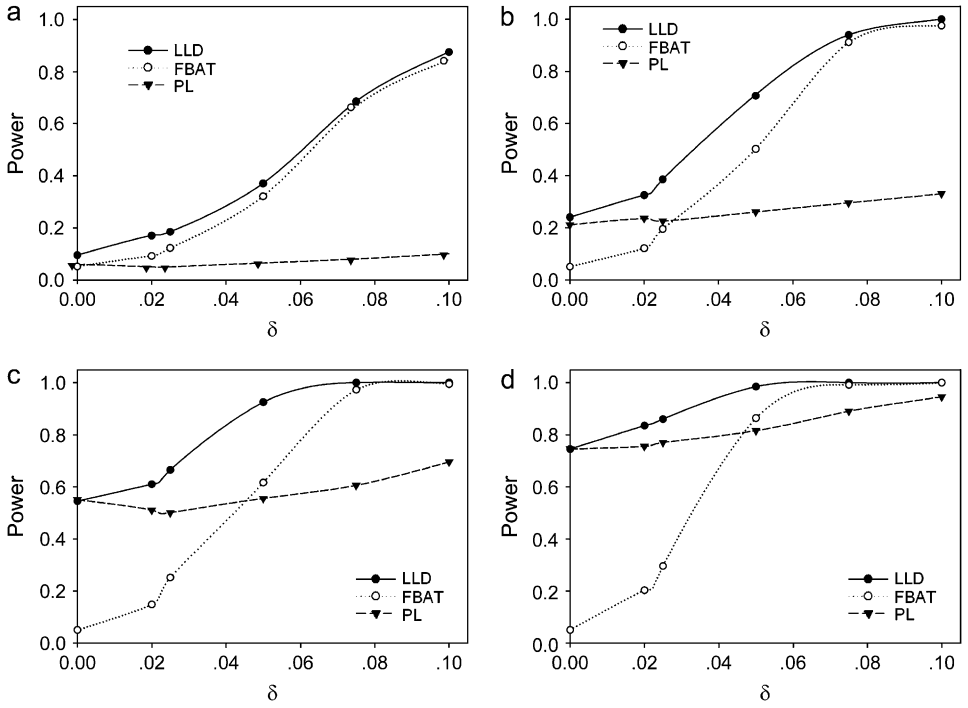

FIGURE 1.—Power of FBAT, PL, and LLD plotted against LD coefficient δ (δ′) at the 0.05 significance level for four heritabilities: (a) $h^2 = 0.1$, (b) $h^2 = 0.2$, (c) $h^2 = 0.3$, and (d) $h^2 = 0.4$, respectively.

power of the FBAT approaches that of LLD. This is not surprising, because more than one sibling is available in our simulations and hence, in theory, information on allele sharing between siblings should contribute to detecting linkage except for the case of no linkage, where there is no practical importance as there is no interest in testing for linkage with a type I error. The power comparison indicates that the union of two complementary components of information allows LLD to be more powerful.

Unlike FBATs, our new approach can also achieve parameter estimation for gene effects, allele frequencies, LD coefficient, and recombination fractions, so that it can provide more knowledge regarding disease etiology. Table 4 lists some typical results on the comparison between LLD and PL, but the results are not shown when LD is absent or weak, as LLD has an estimated result similar to that of PL. In the latter situation, both LLD and PL gave unbiased estimates, although LLD appeared to have slightly larger MSE, but the difference was very small. Such results are highly consistent with our expectation because the assumption of linkage equilibrium is indeed satisfied for linkage-only analysis while one needs to estimate one more unknown

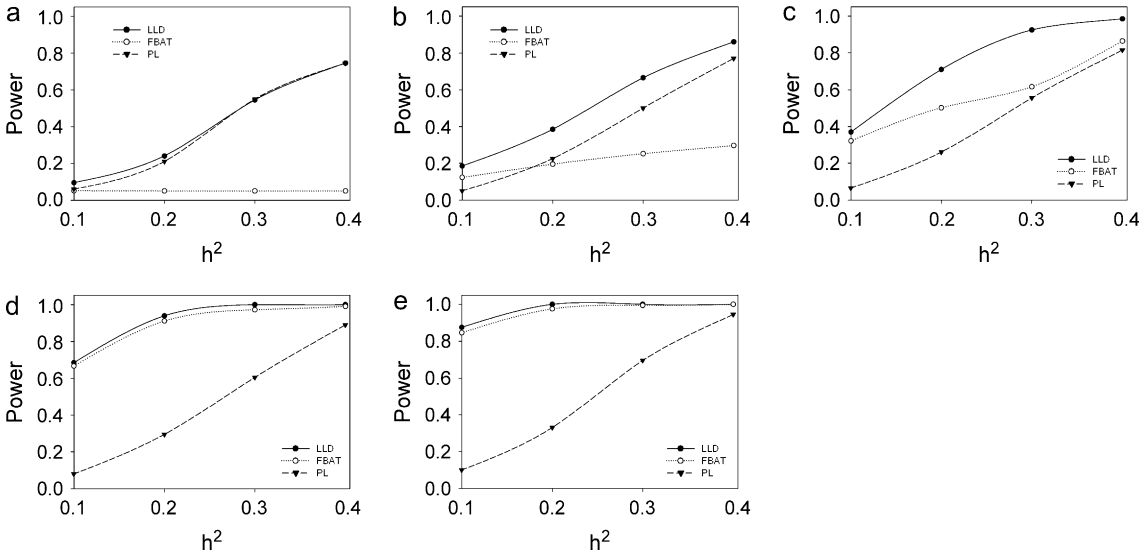

FIGURE 2.—Power of FBAT, PL, and LLD plotted against heritability (h^2) at the 0.05 significance level for five LD coefficients: (a) δ = 0 (δ′ = 0), (b) δ = 0.025 (δ′ = 0.104), (c) δ = 0.05 (δ′ = 0.208), (d) δ = 0.075 (δ′ = 0.313), and (e) δ = 0.1 (δ′ = 0.418), respectively.

TABLE 4

Average MLEs (and root MSEs) from the approach of either pure linkage (PL) or combined linkage and association analysis (LLD)

		LLD		PL	
Parameter	True value	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.1$
$h^2 = 0.2$					
p_Q	0.6	0.549 (0.143)	0.571 (0.135)	0.532 (0.148)	0.545 (0.142)
δ	—	0.059 (0.024)	0.107 (0.026)	—	—
θ	0	0.084 (0.138)	0.049 (0.083)	0.115 (0.197)	0.077 (0.151)
a	0.722	0.661 (0.212)	0.682 (0.164)	0.664 (0.239)	0.675 (0.223)
d	0	0.090 (0.445)	0.013 (0.292)	0.187 (0.580)	0.162 (0.544)
μ	0	0.022 (0.244)	0.035 (0.171)	−0.002 (0.338)	−0.013 (0.311)
σ^2	1	0.971 (0.093)	0.986 (0.088)	0.952 (0.106)	0.953 (0.103)
$h^2 = 0.3$					
p_Q	0.6	0.563 (0.120)	0.585 (0.102)	0.549 (0.122)	0.552 (0.122)
δ	—	0.054 (0.016)	0.104 (0.018)	—	—
θ	0	0.057 (0.098)	0.037 (0.063)	0.066 (0.125)	0.033 (0.074)
a	0.945	0.903 (0.186)	0.928 (0.141)	0.910 (0.183)	0.920 (0.201)
d	0	0.051 (0.349)	0.005 (0.230)	0.101 (0.379)	0.089 (0.370)
μ	0	0.042 (0.190)	0.024 (0.145)	0.039 (0.218)	0.040 (0.221)
σ^2	1	0.980 (0.093)	0.985 (0.097)	0.970 (0.094)	0.962 (0.099)

parameter for LLD. But for cases with slightly stronger LDs, such as $\delta' \geq 0.208$, LLD gained much improvement in estimation accuracy, which is reflected by bias and MSE, over linkage-only analysis (see Table 4). The bias and MSE of the estimated parameters from LLD are almost uniformly less than their counterparts from PL. This is in good agreement with theoretical expectations. In general, the estimation accuracy increases with LD, and LLD improves more than PL. The magnitude of improvement differs with the various parameter values. The estimates of recombination fraction and genetic effects are greatly affected by LD level, while those of population mean and variance are less affected. In some cases, ignoring LD may result in a large bias and MSE in PL. The comparison of parameter estimation strongly indicates that it is necessary to capitalize on the information from population association to get a better and more reliable estimation.

APPLICATION

To demonstrate its use, we applied our new algorithm to a BMD genetic study conducted at Creighton University. A total of 1873 subjects from 405 Caucasian pedigrees containing 740 parents/grandparents and 434 sibships were included in the study. The pedigrees varied in size from 3 to 12 and the mean size was 4.86 while the sibships ranged from 1 to 10 and averaged 2.61. Three SNPs within the vitamin D (1,25-dihydroxyvitamin D3) receptor (VDR) gene, ss12568610, ss12568583, and ss12568608, were chosen to test association with BMD. A detailed description of the clinical subjects and SNP-related information, such as primers/probes and genotyping conditions, has been reported in a separate study (Liu *et al.* 2005). Several BMD-related traits were measured in the study (Liu *et al.* 2005) and we used the BMD at the one-third region of the wrist as an example here. The phenotypic values of the BMD range from 0.349 to 0.997. Our segregation analysis suggested that there is a major gene underlying this trait (data not shown). The coefficients of skewness and kurtosis of the residual effects are 0.123 and 3.175, respectively, which can be regarded as having an approximately normal distribution.

The results analyzed by FBAT and our LLD approach are presented in Tables 5 and 6 for *P*-values, MLEs, and standard errors (SEs), respectively. The three *P*-values in Table 5 are for null hypotheses $\theta = 0.5$, $\theta \geq 0.2$, and $\theta \geq 0.1$, respectively, in LLD analysis, while the *P*-values are for $\theta = 0.5$ in FBAT. After correction for multiple testing, the LR statistic still remains highly significant (minimum $P = 0.001$ for H_0: $\theta = 0.5$), whereas the FBAT statistic shows only marginal significance ($P = 0.040$) for ss12568583. Furthermore, the results of parameter estimation show that all the three SNPs are very tightly linked to the putative disease gene, *i.e.*, have near zero estimated recombination fractions and small SEs, but different frequencies from those of this gene (see Table 6). All estimates from the three SNPs are very consistent, which indicates that, very likely, a gene responsible for BMD is located within or near the VDR gene but the genotyped SNPs do not seem to be the causal variant. The MLEs of δ' are 0.045, 0.177, and 0.021 for SNPs ss12568610, ss12568583, and ss12568608, respectively, suggesting that the associations between the gene and the SNPs are weak. This may be the reason why this gene can elude most FBAT gene-hunting strategies such as QTDT and FBAT. Our approach also gave estimates of

TABLE 5

A comparison of *P*-values for association of VDR SNPs with BMD

SNP	Physical position	Domain	Allele[a]	Allele frequency[b]	FBAT *P*-value	LLD *P*-value		
						$\theta = 0.5$	$\theta \geq 0.2$	$\theta \geq 0.1$
ss12568610	45,470,003	Intron 8	G/**A**	0.419	0.108	0.082	0.123	0.199
ss12568583	45,507,963	5′-UTR	G/**A**	0.280	0.040	0.001	0.021	0.081
ss12568608	45,468,924	Exon 9	T/**C**	0.408	0.246	0.102	0.158	0.241

[a] The boldface type in SNP polymorphisms represents minor alleles.
[b] The allele frequencies are for minor alleles.

the penetrance parameters. As shown in Table 6, the gene has a large genetic effect and displays an incompletely dominant mode of inheritance. In summary, our results indicate that the VDR gene is significantly linked to that for BMD, especially for SNP ss12568583.

DISCUSSION

This study was motivated by the fact that traditional mapping methods, *e.g.*, FBATs and the allele-sharing method, utilize only one component of genetic information, either on linkage or on association, often leading to inefficiency and inaccuracy, although they have desirable properties in some specific cases, *e.g.*, if the assumption of no LD is approximately satisfied in linkage analysis or if the marker tested is exactly the trait gene itself in FBATs. Owing to its ignoring association, the weakness of linkage analysis motivated Risch and Merikangas (1996) to conclude that the allele-sharing method may be hardly up to the task of identifying genes underlying complex traits. On the other hand, however, the TDT may be not as ideal as Risch and Merikangas (1996) claimed because such a perfect case (*i.e.*, perfectly associated, with no recombination and the same allele frequencies) rarely occurs in real data sets, even in a fine-mapping context. Mostly, there are less extreme cases with diverse degrees of LD between both tightly linked loci and loosely linked loci, arising from mutation, recombination erosion, or population admixture. Therefore, it is necessary to develop new approaches that can improve the FBAT's power for successful gene hunting. Heuristically, exploiting information on allele sharing contained in each sibship can achieve this aim and also circumvent the weakness that association is required for linkage to be detected. Such attempts have been pursued by a number of researchers (*e.g.*, Zhao *et al.* 1998; Xiong and Jin 2000; Cantor *et al.* 2005; Li *et al.* 2005). In our previous study (Lou *et al.* 2005), we reported the development of a statistical framework with two hierarchies. In this article, we address a further issue, *i.e.*, developing mapping models with an arbitrary number of hierarchies to handle complex pedigrees. Thus, the proposed approach not only is capable of accommodating multiple loci and/or multiple alleles so that it is easy to tackle interval mapping, multiple interval mapping, and epistasis models, but also allows for diverse types of traits and pedigree structures. Haplotypes in founders contribute information for an association study while informative and partially informative meioses do so for linkage analysis. We unify segregation, linkage, and association analyses into a comprehensive mapping strategy and thus can capture the two complementary aspects of the genetic architecture. The proposed approach has the properties of both linkage analysis and association analysis. From the viewpoint of linkage, it is a LOD score method that is adaptive to the amount of LD. It can make use of LD, if it is indeed present, while it reduces to the standard LOD method when LD is weak or absent. From another viewpoint, it is an association study that incorporates haplotyping analysis in pedigrees and genotyping by a progeny test at a disease locus. For singleton data, it reduces to a parametric association study (*e.g.*, Lou *et al.* 2003; Shibata *et al.* 2004). Although the EM algorithm rather than the quasi-Newton method is used to maximize the likelihood function, our approach is a direct generalization of that of Cantor *et al.* (2005) to multiple loci. The model of Li *et al.* (2005) is also a specific application of the new method, which assumes that the candidate SNP is completely linked to the disease locus and that flanking markers are in linkage equilibrium with one another, the SNP, and the disease

TABLE 6

MLEs (and standard errors) for allele frequency, LD, recombination fraction, additive and dominance effects, mean, and variance on using different VDR SNPs

Parameter	ss12568610	ss12568583	ss12568608
p_Q	0.811 (0.0234)	0.812 (0.0229)	0.812 (0.0234)
δ	−0.004 (0.0066)	0.024 (0.0074)	0.002 (0.0060)
θ	0.000 (0.0025)	0.000 (0.0004)	0.000 (0.0040)
a	0.132 (0.0100)	0.134 (0.0110)	0.132 (0.0100)
d	0.072 (0.0112)	0.070 (0.0110)	0.073 (0.0112)
μ	0.630 (0.0114)	0.630 (0.0115)	0.630 (0.0113)
σ^2	0.006 (0.0002)	0.006 (0.0002)	0.006 (0.0002)

locus. Both the corresponding LR statistics, testing for linkage equilibrium and complete LD, can be constructed by using the new approach.

Another contribution of this article is that it shows, through systematic simulation studies and an application, the important conclusion that a mapping bonus can be obtained by combined linkage and association analysis without any increase in experimental expense. This is highly consistent with theoretical expectation. First, the improvement in mapping resolution arises from the marriage of linkage and association. In a gene-hunting context, latent data exist such as disease genotype, inheritance vector, and linkage phase. Parameter estimation and statistical inference rely on accurate genetic reconstruction of such ambiguous data, *i.e.*, statistical imputation. Violation of the assumption of linkage equilibrium leads to inaccurate imputation in pure linkage analysis so that it may give a biased result, as demonstrated in this article. On the other hand, the assumption of linkage equilibrium also affects imputation precision, owing to its resulting in a likelihood that retains maximum uncertainty about which component of the mixture distribution generates the data, and hence is least informative for the recombination parameter θ. Theoretically, integrating both complementary components increases imputation accuracy, leading to improvement in mapping accuracy, precision, and power over traditional linkage analysis. Intuitively, linkage induces more gene concordance between related individuals having similar phenotypes, while the opposite holds true for those with disparate phenotypes. Incorporating this information, which FBATs fail to do, gives our LLD approach a higher power than that of FBATs. This type of phenomenon has also been widely observed with comparisons of the TDT, the conventional affected sib pairs, and combined methods (Huang and Jiang 1999; Wicks and Wilson 2000; Lazzeroni 2002). Both our simulation and real data studies support our theoretical expectation.

Second, the improvement may also come from two other potential sources, although they are not explored in this article. The LLD approach integrates population-based association analysis and pedigree-based linkage analysis into a coherent framework so that it can handle diverse types of data, including full sibs, half sibs, cousins, nuclear families, extended nuclear families, complex pedigrees, and singletons, as well as their mixtures. Unlike those of Huang and Jiang (1999), Wicks and Wilson (2000), and Lazzeroni (2002), which require only affected pairs with at least one heterozygous parent, our approach allows for analyzing any type of data structure, including singletons and pedigrees without any informative meioses, which do not contribute information to linkage parameter(s) but do inform association parameter(s). Without dropping any type of mapping data, we make use of data to the maximum extent, leading to the possibility of improving mapping performance. Furthermore, our flexible framework is easily applied to a multipoint analysis. It has been well documented that multipoint analyses can extract more statistical information than pairwise ones and thus may substantially increase the power and reduce spurious results (Lathrop *et al.* 1984, 1985). Conceivably, unifying multilocus linkage and association mapping will further improve mapping resolution.

Our current version of the program is capable of handling five to six loci on a PC computer if only limited amounts of data are missing. Although it allows for more loci and alleles on a workstation or a PC cluster with more memory and storage, computation can be very time-consuming for a large number of loci and alleles because the required memory and time exponentially increase with the number of loci. To avoid a formidable computational burden, the simulation-based versions of the EM algorithm such as stochastic EM and Markov chain Monte Carlo (MCMC) EM (Thompson 1994; Celeux *et al.* 1996), based on *estimating* conditional posterior probabilities in the E-step rather than computing them exactly, can be used. The approximate methods based a composite likelihood (*e.g.*, Rannala and Slatkin 2000) also seem to be the feasible ways to tackle this problem. The relevant study is under way.

Moreover, unlike the model-free approaches such as allele sharing and FBATs, which can tell us only whether linkage or association exists but fail to provide any estimates of what values they have, the proposed LLD approach can simultaneously provide parameter estimation of genetic distance, allelic association, and genotype–phenotype relationship and also perform various types of hypothesis testing. For example, we can perform a comparison of the analyses, including or not the LD information to assess the validity of the LD model assumptions. Thus, this approach exposes more genetic mechanisms than FBATs to genetic etiology and hence increases the predictability of gene mapping.

Finally, as pointed out by Pérez-Enciso (2003), given the diversity of genetic architectures and population histories, it is unlikely that a single statistical approach will be valid for all cases. The approach described here is subject to the same limitations faced by all model-based methods, *i.e.*, the requirement of a correct, or close to correct, model for the trait under study. If the model for predicting disease status from phenotypes is not sufficiently well known, this approach cannot perform well. Therefore, this model-based LLD approach should serve as a supplement to model-free methods in tracking the gene(s) underlying complex diseases, once model-free methods have suggested how many loci are involved and their approximate locations in the genome (Elston 1998).

This work is funded in part by a grant from the National Institute on Drug Abuse to M. D. Li (DA-12844), by grants from the National Center for Research Resources (RR03655) and the National Institute of General Medical Sciences (GM28356) to R. C. Elston, and by a grant from the National Science Foundation of China (30000097) to X.-Y. Lou.

LITERATURE CITED

Abecasis, G. R., L. R. Cardon and W. O. Cookson, 2000 A general test of association for quantitative traits in nuclear families. Am. J. Hum. Genet. **66:** 279–292.

Abecasis, G. R., S. S. Cherny, W. O. Cookson and L. R. Cardon, 2002 Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. **30:** 97–101.

Allison, D. B., M. Heo, N. Kaplan and E. R. Martin, 1999 Sibling-based tests of linkage and association for quantitative traits. Am. J. Hum. Genet. **64:** 1754–1763.

Amos, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. Am. J. Hum. Genet. **54:** 535–543.

Cantor, R. M., G. K. Chen, P. Pajukanta and K. Lange, 2005 Association testing in a linked region using large pedigrees. Am. J. Hum. Genet. **76:** 538–542.

Celeux, G., D. Chauveau and J. Diebolt, 1996 Stochastic versions of the EM algorithm: an experimental study in the mixture case. J. Stat. Comput. Simul. **55:** 287–314.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39:** 1–38.

Elston, R. C., 1998 Linkage and association. Genet. Epidemiol. **15:** 565–576.

Elston, R. C., and J. Stewart, 1971 A general model for the genetic analysis of pedigree data. Hum. Hered. **21:** 523–542.

Falk, C. T., and P. Rubinstein, 1987 Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann. Hum. Genet. **51**(3): 227–233.

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.

Gudbjartsson, D. F., K. Jonasson, M. L. Frigge and A. Kong, 2000 Allegro, a new computer program for multipoint linkage analysis. Nat. Genet. **25:** 12–13.

Haseman, J. K., and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2:** 3–19.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin *et al.*, 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072–1079.

Huang, J., and Y. Jiang, 1999 Linkage detection adaptive to linkage disequilibrium: the disequilibrium maximum-likelihood-binomial test for affected-sibship data. Am. J. Hum. Genet. **65:** 1741–1759.

International HapMap Consortium, 2003 The International HapMap Project. Nature **426:** 789–796.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

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

Kruglyak, L., M. J. Daly, M. P. Reeve-Daly and E. S. Lander, 1996 Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. **58:** 1347–1363.

Laird, N. M., S. Horvath and X. Xu, 2000 Implementing a unified approach to family-based tests of association. Genet. Epidemiol. **19**(Suppl. 1): S36–S42.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

Lange, K., and R. C. Elston, 1975 Extensions to pedigree analysis I. Likelihood calculations for simple and complex pedigrees. Hum. Hered. **25:** 95–105.

Lange, C., and N. M. Laird, 2002 Power calculations for a general class of family-based association tests: dichotomous traits. Am. J. Hum. Genet. **71:** 575–584.

Lange, C., D. L. DeMeo and N. M. Laird, 2002 Power and design considerations for a general class of family-based association tests: quantitative traits. Am. J. Hum. Genet. **71:** 1330–1341.

Lathrop, G. M., and J. M. Lalouel, 1984 Easy calculations of lod scores and genetic risks on small computers. Am. J. Hum. Genet. **36:** 460–465.

Lathrop, G. M., J. M. Lalouel, C. Julier and J. Ott, 1984 Strategies for multilocus linkage analysis in humans. Proc. Natl. Acad. Sci. USA **81:** 3443–3446.

Lathrop, G. M., J. M. Lalouel, C. Julier and J. Ott, 1985 Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am. J. Hum. Genet. **37:** 482–498.

Lazzeroni, L. C., 2002 Allele sharing and allelic association I: sib pair tests with increased power. Genet. Epidemiol. **22:** 328–344.

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

Li, M., M. Boehnke and G. R. Abecasis, 2005 Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. Am. J. Hum. Genet. **76:** 934–949.

Liu, P. Y., Y. Y. Zhang, Y. Lu, J. R. Long, H. Shen *et al.*, 2005 A survey of haplotype variants at several disease candidate genes: the importance of rare variants for complex diseases. J. Med. Genet. **42:** 221–227.

Lou, X.-Y., G. Casella, R. C. Littell, M. C. Yang, J. A. Johnson *et al.*, 2003 A haplotype-based algorithm for multilocus linkage disequilibrium mapping of quantitative trait loci with epistasis. Genetics **163:** 1533–1548.

Lou, X.-Y., G. Casella, R. J. Todhnuter, M. Yang and R. Wu, 2005 A general statistical framework for unifying interval and linkage disequilibrium mapping: towards high-resolution mapping of quantitative traits. J. Am. Stat. Assoc. **100:** 158–171.

Louis, T. A., 1982 Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B **44:** 226–233.

MacLean, C. J., N. E. Morton and S. Yee, 1984 Combined analysis of genetic segregation and linkage under an oligogenic model. Comput. Biomed. Res. **17:** 471–480.

Meng, X.-L., and D. B. Rubin, 1991 Using EM to obtain asymptotic variance-convariance matrices: the SEM algorithm. J. Am. Stat. Assoc. **86:** 899–909.

O'Connell, J. R., and D. E. Weeks, 1995 The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat. Genet. **11:** 402–408.

Ott, J., 1974 Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am. J. Hum. Genet. **26:** 588–597.

Pérez-Enciso, M., 2003 Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. Genetics **163:** 1497–1510.

Rabinowitz, D., and N. Laird, 2000 A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum. Hered. **50:** 211–223.

Rannala, B., and M. Slatkin, 2000 Methods for multipoint disease mapping using linkage disequilibrium. Genet. Epidemiol. **19**(Suppl. l): S71–S77.

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

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

Shibata, K., T. Ito, Y. Kitamura, N. Iwasaki, H. Tanaka *et al.*, 2004 Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics **168:** 525–539.

Spielman, R. S., R. E. McGinnis and W. J. Ewens, 1993 Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. **52:** 506–516.

Thompson, E. A., 1994 Monte Carlo likelihood in genetic mapping. Stat. Sci. **9:** 355–366.

Ward, P. J., 1993 Some developments on the affected-pedigree-member method of linkage analysis. Am. J. Hum. Genet. **52:** 1200–1215.

Whittaker, J. C., M. C. Denham and A. P. Morris, 2000 The problems of using the transmission/disequilibrium test to infer tight linkage. Am. J. Hum. Genet. **67:** 523–526.

Wicks, J., 2000 Exploiting excess sharing: a more powerful test of linkage for affected sib pairs than the transmission/disequilibrium test. Am. J. Hum. Genet. **66:** 2005–2008.

Wicks, J., and S. R. Wilson, 2000 Evaluating linkage and linkage disequilibrium: use of excess sharing and transmission disequilibrium methods in affected sib pairs. Ann. Hum. Genet. **64:** 419–432.

Xiong, M., and L. Jin, 2000 Combined linkage and linkage disequilibrium mapping for genome screens. Genet. Epidemiol. **19:** 211–234.

Zhao, L. P., C. Aragaki, L. Hsu and F. Quiaoit, 1998 Mapping of complex traits by single-nucleotide polymorphisms. Am. J. Hum. Genet. **63:** 225–240.

Communicating editor: C. Haley

APPENDIX

For affected sib pair data with a recessively inherited disease as described in the text ($p = 0.5$), under the assumption of linkage equilibrium we have the log-likelihood

$$\begin{aligned}\ln L_A = \text{Constant} &+ (2n_{AA,AA} + 2n_{aa,aa} + n_{AA,Aa} + n_{Aa,aa})\ln[\theta^2 + (1-\theta)^2 + 4] \\ &+ (2n_{AA,aa} + n_{AA,Aa} + n_{Aa,aa})\ln[2\theta(1-\theta) + 1] + n_{Aa,Aa}\ln\{[\theta^2 + (1-\theta)^2 + 4]^2 + [2\theta(1-\theta) + 1]^2\}, \quad \text{(A1)}\end{aligned}$$

where $n_{AA,AA}$, $n_{AA,Aa}$, $n_{AA,aa}$, $n_{Aa,Aa}$, $n_{Aa,aa}$, and $n_{aa,aa}$ are the numbers of affected sib pairs with marker configurations $\{AA, AA\}$, $\{AA, Aa\}$, $\{AA, aa\}$, $\{Aa, Aa\}$, $\{Aa, aa\}$, and $\{aa, aa\}$, with true probabilities $[(1-\theta_0)^2 + 2]^2/9$, $4\theta_0(1-\theta_0)[(1-\theta_0)^2 + 2]/9$, $2\theta_0^2(1-\theta_0)^2/9$, $4\theta_0^2[(1-\theta_0)^2 + 1]/9$, $4\theta_0^3(1-\theta_0)/9$, and $\theta_0^4/9$, respectively; and θ and θ_0 (a specific value) are the recombination fractions between the marker and the disease gene. The partial derivative with respect to θ, the score, is

$$\begin{aligned}\frac{\partial \ln L_A}{\partial \theta} &= 2(1-2\theta)\left\{\frac{2n_{AA,aa} + n_{AA,Aa} + n_{Aa,aa}}{2\theta(1-\theta) + 1} - \frac{2n_{AA,AA} + 2n_{aa,aa} + n_{AA,Aa} + n_{Aa,aa}}{\theta^2 + (1-\theta)^2 + 4} - \frac{4n_{Aa,Aa}[\theta^2 + (1-\theta)^2 + 1]}{[\theta^2 + (1-\theta)^2 + 4]^2 + [2\theta(1-\theta) + 1]^2}\right\} \\ &= 2(1-2\theta)\left\{\frac{2(n_{AA,Aa} + n_{Aa,aa}) + 8n_{AA,aa} - 4(n_{AA,AA} + n_{aa,aa}) - 2[\theta^2 + (1-\theta)^2](n_{AA,AA} + n_{aa,aa} - n_{AA,aa})}{[2\theta(1-\theta) + 1][\theta^2 + (1-\theta)^2 + 4]}\right\} \\ &\quad - \frac{8n_{Aa,Aa}(1-2\theta)[\theta^2 + (1-\theta)^2 + 1]}{[\theta^2 + (1-\theta)^2 + 4]^2 + [2\theta(1-\theta) + 1]^2}. \quad \text{(A2)}\end{aligned}$$

Because

$$\begin{aligned}&2(n_{AA,Aa} + n_{Aa,aa}) + 8n_{AA,aa} - 4(n_{AA,AA} + n_{aa,aa}) - 2[\theta^2 + (1-\theta)^2](n_{AA,AA} + n_{aa,aa} - n_{AA,aa}) \\ &\quad \approx -n\left\{4\frac{(1-\theta_0)^4 + \theta_0^4 + 1 + 3\theta_0^2 + 7(1-\theta_0)^2}{9} + 2[\theta^2 + (1-\theta)^2]\frac{(1-2\theta_0)^2 + 4(1-\theta_0)^2 + 4}{9}\right\} < 0,\end{aligned}$$

whatever value θ takes, the assumed likelihood is a monotonically decreasing function of θ in the interval $[0, 0.5]$, and hence $\hat{\theta} = 0$ is the MLE, even if there is no linkage; *i.e.*, $\theta_0 = 0.5$.