Original Paper

Robust Multipoint Simultaneous Identical-by-Descent Mapping for Two Linked Loci

Wan-Yu Lin
Institute of Epidemiology
National Taiwan University
Taipei, Taiwan

Daniel J. Schaid [1]
Division of Biostatistics
Mayo Clinic
Rochester, MN

November 6, 2006

[1] Correspondence: Dr. Daniel J. Schaid, Division of Biostatistics, Harwick 7, Mayo Clinic, 200 First Street SW, Rochester, MN 55905 (USA). Tel. +1 507 284 0639, Fax +1 507 284 9542, E-mail: schaid@mayo.edu

Abstract

A challenging issue in genetic mapping of complex human diseases is localizing disease susceptibility genes when the genetic effects are small to moderate. There are greater complexities when multiple loci are linked to a chromosomal region. Liang et al. [1] proposed a robust multipoint method that can simultaneously estimate both the position of a trait locus and its effect on disease status by using affected sib pairs (ASPs). Based on the framework of generalized estimating equations (GEEs), the estimate and standard error of the position of a trait locus are robust to different genetic models. To utilize other relative pairs collected in pedigree data, Schaid et al. [2] extended Liang's method to various types of affected relative pairs (ARPs) by two approaches: unconstrained and constrained methods. However, the above methods are limited to situations in which only one trait locus exists on the chromosome of interest. The mean functions are no longer correctly specified when there are multiple causative loci linked to a chromosomal region. To overcome this, Biernacka et al. [3] considered the multipoint methods for ASPs to allow for two linked disease genes. We further generalize the approach to cover other types of ARPs. To reflect realistic situations for complex human diseases, we set modest sizes of genetic effects in our simulation. Our results suggest that several hundred independent pedigrees are needed, and markers with high information, to provide reliable estimates of trait locus positions and their confidence intervals. Bootstrap resampling can correct the downward bias of the robust variance for location estimates. These methods are applied to a prostate cancer linkage study on chromosome 20 and compared with the results for the one-locus model [2]. We have implemented the multipoint IBD mapping for one and two linked loci in our software GEEARP, which allows analyses for five general types of ARPs.

Key Words:

Affected relative pairs; Affected sib pairs; Generalized estimating equations; Linkage analysis.

1 Introduction

Most complex human diseases are likely induced by more than one disease susceptibility gene, and the genetic effects for those genes are usually small to moderate. There are two approaches to map genes, linkage and association analyses. The former is based on familial data. The power comes from distinguishing recombinant/nonrecombinant chromosomal segments transmitted from parents to offspring, and usually requires a large number of families to find evidence for linkage [4]. Association analyses are for fine mapping, which can be population-based or family-based. Here we focus on linkage studies and develop a robust multipoint mapping method for two linked causative loci. The model-free approach that we use does not need a prespecified genetic model, and relies on the identity-by-descent (IBD) sharing of affected relative pairs (ARPs). Liang et al. [1] proposed an IBD-based procedure, on the basis of affected sib pairs (ASPs), to estimate the location of an unobserved susceptibility gene within a chromosomal region framed by multiple markers. They derived the expected IBD scores on markers for ASPs, which are functions of the distance from the disease gene and the genetic effect. Further, they introduced a generalized estimating equation (GEE) approach to estimate the location and genetic effect simultaneously, as well as confidence intervals based on a robust variance estimator. An advantage of GEE [5] is that it does not require a presumption of the underlying genetic model, and it is robust to a wide variety of genetic mechanisms. Schaid et al. [2] extended this method to different types of ARPs by two approaches, unconstrained and constrained. If there is no epistasis and no dominance, their constrained model provides a good solution to reduce the number of parameters, which is essential when there are not many ARPs other than ASPs. Without prior knowledge of genetic mechanisms, a score test they provide can be used to examine the appropriateness of fitting the constrained model. The main assumption of the above two papers is that there is only one trait locus on the chromosome of interest. This might not be the situation for some complex human diseases. To jointly localize two linked disease loci, Biernacka et al. [3] derived an expression for expected allele sharing among ASPs on a chromosomal segment containing two disease susceptibility genes. Their simulations covered several important issues, including the influence of genetic models, sample size, distance between two disease loci, marker properties (i.e., densities, numbers, and informativeness). In another paper, they provided tests for the presence of two linked disease susceptibility genes, including a quasi-likelihood ratio test and a modified score test [6]. A preliminary analysis by Kong-Cox LOD scores (KC-LOD) [7] or by the approach of Biernacka et al. can be used to judge whether two-locus model should be adopted. Their work is based on ASP data, which may be a convenient sampling unit, but means that one discards information from other ARPs when they are available. Here we generalize the two-locus localization method to a variety of ARPs. In the Methods section, we develop the approach to account for multipoint mapping for two loci simultaneously, using multiple types of ARPs. We describe it for both an unconstrained model and a constrained model [2], along with a score test generalized from Schaid et al. [2] to examine the appropriateness of fitting a constrained model. In the Simulation section, by setting two-locus additive models and two-locus multiplicative models with modest genetic effects (or modest recurrence risk ratios), we evaluate

the methods by our own simulation programs. Finally, an application to prostate cancer linkage on chromosome 20 is illustrated.

2 Methods

2.1 Unconstrained-Model Approach

Consider that there are two linked disease susceptibility genes located at positions τ_1 and τ_2 in a chromosomal region, as shown in Figure 1. The expected number of alleles IBD for the linked markers will be greater than that for the unlinked markers for an ARP. Let $f_{ji}(t)$ be the probability that the ith ARP has j alleles IBD at location t $(j = 0, 1, 2)$. The IBD score for the ith ARP is estimated by

$$S_i(t) = 2f_{2i}(t) + f_{1i}(t).$$

Let θ_1 and θ_2 be the recombination fractions between a marker at location t and the disease genes at positions τ_1, τ_2, respectively. Let θ_3 be the recombination fraction between the two disease genes. We use Haldane's mapping function [8] to relate the recombination fraction to the genetic distance,

$$\theta_1 = (1 - e^{-0.02|\tau_1 - t|})/2,$$

$$\theta_2 = (1 - e^{-0.02|\tau_2 - t|})/2,$$

$$\theta_3 = (1 - e^{-0.02|\tau_2 - \tau_1|})/2,$$

where t, τ_1, τ_2 are in centiMorgans (cM). Under the assumptions of random mating, linkage equilibrium, generalized single ascertainment [1, 3, 9], no interference, and equal recombination fractions for males and females, the expected IBD scores at location t for five types of ARPs are summarized in Table 1. The derivation depends on the joint distribution of IBD and follows Biernacka et al. [3] for ASPs. The IBD scores at disease genes for ARPs tend to be greater than random sharing, so do those at the adjacent markers. The two disease genes separate the chromosomal region into three intervals. When $t < \tau_1 < \tau_2$, under the assumption of no interference, $Pr(S(t) = j|S(\tau_1) = l, S(\tau_2) = m) = Pr(S(t) = j|S(\tau_1) = l)$, the first-order Markov property simplifies the mean function of alleles shared IBD at location t to be dependent on only the first disease gene. When $\tau_1 < \tau_2 < t$, under the same assumption, $Pr(S(t) = j|S(\tau_1) = l, S(\tau_2) = m) = Pr(S(t) = j|S(\tau_2) = m)$, the expected IBD scores at location t depends only on the second disease gene. To sum up, the mean IBD sharing on markers located in the two intervals is the same as that derived for a one-locus model [1, 2]. With the assumption of no interference, the mean function that differs from the framework of the one-locus model is only when $\tau_1 < t < \tau_2$, in which case the expected alleles shared IBD at t depends on both disease genes. The expected IBD score in the middle interval is higher than that considering only one gene in that region. Omitting the second disease susceptibility gene would bias the localization for the first gene. The unknown parameters in the mean function are locations (τ_1, τ_2) and genetic effects (C_{1k} and C_{2k}, the effects of genes at τ_1 and τ_2 for the kth type of ARP). Note that in the situation of two linked disease susceptibility genes, C_{1k} and C_{2k} do not represent the marginal effects of gene 1 and gene 2. Rather, they incorporate the gene effect of each other, and thus they change with the recombination fraction between them. As θ_3 gets smaller, a C coefficient combines more of the genetic effect of the other locus, and thus gets larger. We also condition on Φ, the event of an ARP. We let $k = 1, \cdots, 5$ denote full siblings (FS), half siblings (HS), first cousins

(FC), grandparent-grandchild pairs (GP), and avuncular pairs (AP), respectively. Let $\boldsymbol{\gamma}$ be the vector composed of the twelve parameters,

$$\boldsymbol{\gamma} = (\tau_1, \tau_2, C_{11}, C_{21}, C_{12}, C_{22}, C_{13}, C_{23}, C_{14}, C_{24}, C_{15}, C_{25}).$$

Assume there are N ARPs and M markers genotyped in a selected chromosomal region, $\boldsymbol{S}_i$ is the IBD-score vector for the ith ARP on the M markers, with mean vector $E(\boldsymbol{S}_i|\Phi)$ and variance-covariance matrix $Cov(\boldsymbol{S}_i|\Phi)$. Let $\boldsymbol{U}_i$ be the score vector provided by the ith ARP, which is

$$\boldsymbol{U}_i = \left[\frac{\partial E(\boldsymbol{S}_i|\Phi)}{\partial \boldsymbol{\gamma}}\right]^T Cov^{-1}(\boldsymbol{S}_i|\Phi)[\boldsymbol{S}_i - E(\boldsymbol{S}_i|\Phi)]. \tag{1}$$

Summing over all the N ARPs and solving the system of estimating equations,

$$\boldsymbol{U} = \sum_{i=1}^{N} \left[\frac{\partial E(\boldsymbol{S}_i|\Phi)}{\partial \boldsymbol{\gamma}}\right]^T Cov^{-1}(\boldsymbol{S}_i|\Phi)[\boldsymbol{S}_i - E(\boldsymbol{S}_i|\Phi)] = \boldsymbol{0}, \tag{2}$$

obtains the point estimate of $\boldsymbol{\gamma}$. In the conventional GEE framework, $Cov(\boldsymbol{S}_i|\Phi)$ is formulated as functions of the mean and the correlation between markers. However, to reduce the computational burden, we use the empirical variances of alleles IBD on the diagonal and let off-diagonal covariances be 0, giving our working covariance matrix, $\boldsymbol{W}$. This independence working covariance matrix avoids over-formulated covariances and the inversion of an ill-conditioned matrix in every iteration of Fisher's scoring method. This leads to a much more convenient method to solve (2). Some critical numerical methods for our algorithm are given in the Appendix.

The precision is assessed by the robust variance,

$$Var(\hat{\boldsymbol{\gamma}}) = \mathbf{I}_w^{-1}\mathbf{I}_r\mathbf{I}_w^{-1}, \tag{3}$$

where $\mathbf{I}_w$ is the working information matrix given by

$$\mathbf{I}_w = \sum_{i=1}^{N}[\frac{\partial E(\boldsymbol{S}_i|\Phi)}{\partial \boldsymbol{\gamma}}]^T \boldsymbol{W}^{-1}[\frac{\partial E(\boldsymbol{S}_i|\Phi)}{\partial \boldsymbol{\gamma}}], \tag{4}$$

and $\mathbf{I}_r$ is the empirical variance,

$$\mathbf{I}_r = \sum_{j=1}^{P} \boldsymbol{U}_j \boldsymbol{U}_j{}^T, \tag{5}$$

in which $\boldsymbol{U}_j$ is the sum of score vectors for all pairs in the jth pedigree ($j = 1, \cdots, P$). The sandwich estimator in (3) accounts for misspecification of the working covariance matrix and any correlation among multiple ARPs from a same pedigree. However, in simulation studies described in Section 3.2, we found that the robust sandwich estimator tends to give smaller standard errors for $\hat{\tau}_1$ and $\hat{\tau}_2$, compared with the simulation-based standard errors. A bootstrap resampling method yields standard errors much closer to the simulation-based standard errors, although this increases the computational intensity.

2.2 Constrained-Model Approach

The unconstrained model estimates common locations τ_1, τ_2, and different genetic effects for each type of ARPs, C_{ik}, $i = 1, 2$, $k = 1, \cdots, 5$. When there are several kinds of ARPs included, the method involves a large number of parameters. This can cause considerable variability in parameter estimates, especially when the numbers of each type of ARP are small. To reduce the number of parameters, we consider the constrained-model approach provided by Schaid et al. [2]. Based on Risch [10], the risk ratio λ is a function of genetic effect C (see [2] or Table 1, here we drop subscripts for clarity). Let λ_{ik} be the risk ratio transformed from C_{ik}, with subscript i denoting disease locus and k denoting type of ARP. Under no dominance and no epistasis, all λ_{ik}'s reduce to a λ_i, $i = 1$ or 2 [10]. That is, the diverse genetic effects C_{ik}'s can be formulated as functions of only one parameter, λ_i, $i = 1$ or 2. In this situation, it is not necessary to use the model mentioned in the previous subsection, and we can reduce it to one that contains four parameters. This simpler model is subjected to specific functions relating genetic effects across types of ARPs, and thus it has common ratios λ_1 and λ_2. λ_i can be explained as the risk ratio for a pair of relatives sharing one allele IBD at the ith trait locus, compared with a pair of relatives sharing no allele IBD at that locus. Thus, when there is no epistasis and no dominance for both loci, we can apply the constrained-model approach to reduce the number of unknown parameters, and to increase the statistical efficiency. Let $\boldsymbol{\gamma_c}$ be the vector composed of the four parameters in the constrained model,

$$\boldsymbol{\gamma_c} = (\tau_1, \tau_2, \lambda_1, \lambda_2).$$

The score equations and robust variance are defined in the same way as in the unconstrained model, (2) and (3), but now the parameter vector is $\boldsymbol{\gamma_c}$.

2.3 Testing Homogeneity of λ_{ik}

If the λ_{1k}'s are all equal and the λ_{2k}'s are all equal (with varying k standing for different types of ARPs), the unconstrained model could be reduced to a constrained model with only four parameters $\boldsymbol{\gamma_c} = (\tau_1, \tau_2, \lambda_1, \lambda_2)$. This holds when there is no dominace nor epistasis for both loci. To evaluate the appropriateness of fitting a constrained model, we generalize the score test provided by Schaid et al. [2] to simultaneously test the homogeneity of λ_{1k}'s and λ_{2k}'s. The null hypothesis is $H_0 : \lambda_{11} = \cdots = \lambda_{1k} = \cdots = \lambda_{1K}, \lambda_{21} = \cdots = \lambda_{2k} = \cdots = \lambda_{2K}$, where K is the number of types of ARPs. For illustration, to test the homogeneity of λ_{ik} between FS, FC, and AP, the matrix $\mathbf{H}$ is

$$\mathbf{H} = \left\{ \begin{array}{cccccccc} 0 & 0 & \frac{\partial\lambda_{11}}{\partial C_{11}} & 0 & -\frac{\partial\lambda_{13}}{\partial C_{13}} & 0 & 0 & 0 \\ 0 & 0 & \frac{\partial\lambda_{11}}{\partial C_{11}} & 0 & 0 & 0 & -\frac{\partial\lambda_{15}}{\partial C_{15}} & 0 \\ 0 & 0 & 0 & \frac{\partial\lambda_{21}}{\partial C_{21}} & 0 & -\frac{\partial\lambda_{23}}{\partial C_{23}} & 0 & 0 \\ 0 & 0 & 0 & \frac{\partial\lambda_{21}}{\partial C_{21}} & 0 & 0 & 0 & -\frac{\partial\lambda_{25}}{\partial C_{25}} \end{array} \right\}. \tag{6}$$

The score statistic to test H_0 is

$$T = \tilde{U}'\tilde{\mathbf{I}}_{\mathbf{w}}^{-1}\tilde{\mathbf{H}}'(\tilde{\mathbf{H}}\tilde{\mathbf{I}}_{\mathbf{w}}^{-1}\tilde{\mathbf{I}}_{\mathbf{r}}\tilde{\mathbf{I}}_{\mathbf{w}}^{-1}\tilde{\mathbf{H}}')^{-1}\tilde{\mathbf{H}}\tilde{\mathbf{I}}_{\mathbf{w}}^{-1}\tilde{U}. \tag{7}$$

The score vector (2), working information matrix (4), and empirical variance (5) are defined as before, but now they are evaluated under the null hypothesis (indicated by tilde). We need to fit the constrained model first, transforming the λ_i estimates into C_{ik} coefficients (see Table 1), putting them into the unconstrained model together with τ_i estimated in the constrained model. The statistic T has an approximate chi-square distribution with degrees of freedom $2(K-1)$.

3 Simulation Study

3.1 Simulation Scenario

Consider a hypothetical chromosomal region with 11 markers and two disease genes as shown in Figure 1. The markers are spaced at 10 cM intervals and the two disease genes are 40 cM apart ($\tau_1 = 35$ cM and $\tau_2 = 75$ cM). We simulated data for 500 independent pedigrees with structure shown in Figure 2. Fully informative marker IBD sharing for ARPs under a variety of two-locus models were generated by our program. Families were included in the study if they had at least two affected members in the third generation. We only considered three common types of ARPs in the simulation: full siblings (FS), first cousins (FC), and avuncular pairs (AP). Under the pedigree structure and the ascertainment scheme, we would generate more affected pairs that are FC and AP than FS. To reflect common ascertainment schemes that oversample affected siblings, we included FC only when subjects 9 and 11 were both affected, but ignored FC composed by other individuals. Similarly, to reduce the number of AP, we included AP composed of subjects 3 and 10, or subjects 4 and 8, but ignored other AP. However, all FS were kept. Thus, any included pedigree had at most one FC, at most two AP, and the number of FS ranged up to five. By this, we had more FS than FC and AP. Each simulation result was based on 1,000 repetitions, except for bootstrap evaluations that were based on 100 repetitions with 500 bootstrap samples for each repetition. Genotype data were generated along chromosomes for the founders of the pedigrees. Assuming random mating, the genotypes on the two disease genes were simulated by the models presented in Table 2. We considered four models with varying penetrance matrices and disease allele frequencies. Let there be two diallelic disease genes, with alleles (A,a) and (B,b), respectively. Denote the genotypes at the first locus by G_i, $i = 1, 2, 3$, with corresponding population frequencies p_i and those at the second locus by H_j, $j = 1, 2, 3$, with corresponding population frequencies q_j. By the definition of Risch [11], the first two models were two-locus additive models in the sense that the penetrance w_{ij} is the sum of penetrance factors x and y, that is, $w_{ij} = x_i + y_j$. The last two models were two-locus multiplicative models in the sense that the penetrance w_{ij} is the product of penetrance factors x and y, that is, $w_{ij} = x_i \times y_j$. The prevalence of disease was $\sum_{i=1}^{3} \sum_{j=1}^{3} p_i q_j w_{ij}$. The additive model is characterized by no interlocus interaction, and was shown to closely approximate genetic heterogeneity [11]. In contrast, the multiplicative model can represent epistasis between loci. These models were chosen to represent a range of genetic models including additivity and epistasis. Under the setting of a two-locus multiplicative model, the genetic effect size of the first disease gene in Model B is very small, which leads Model B to be close to a one-locus model. Note that the four models were all set for complex diseases, so the genetic effects and risk ratios were small to moderate.

3.2 Simulation Results

Tables 3 and 4 illustrate the simulation results for true values of $\tau_1 = 35$ and $\tau_2 = 75$ (cM). The true values for genetic effects C_{ik} and the corresponding transformed λ_{ik} parameters are listed in Table 2. The initial values in the Fisher's scoring method were set at $\tau_1 = 50, \tau_2 =$

$60, C's = 0.1$ for unconstrained models and $\tau_1 = 50, \tau_2 = 60, \lambda's = 1.2$ for constrained models. Results with initial values of $\tau_1 = 20, \tau_2 = 90$ were very similar (not shown). The first location estimate in Model B incorporates large variability and hence a large mean squared error even with 500 pedigrees. This is because the genetic effect size is too small to be detected. From the simulation results for unconstrained models, we find there is a notable downward bias in the variance estimates provided by the robust variance estimator, so the 95% CI coverage is much less than the nominal confidence interval coverage. This only happened for the location estimates, not the genetic effects. The robust variance estimator gives a good approximation to the simulation-based standard error for the estimated genetic effect sizes. In the constrained models, the downward bias in the variance estimates for the location parameters is not as severe, but there is a slight upward bias in the variance estimates for the estimated risk ratios. To overcome the downward bias caused by the robust variance estimator, a bootstrap standard error was obtained from 500 bootstrap samples drawn with replacement from a sample 500 families. This was closer to the empirical standard error obtained from 1,000 simulated samples. Tables 3 and 4 list bootstrap point estimates and standard errors, but only for 100 replicates because of heavy computational burden. The 95% CI was calculated from (2.5-97.5) percentiles of bootstrap estimates. The coverage is close to the nominal level. When the sample size was moderate (say, 100 pedigrees, results not shown), there was a large variability in the parameter estimates. Since the genetic effects are small, and there are multiple parameters to be estimated, this localization method is not expected to perform well for relatively small sample sizes.

From Table 3, except FC, there is a general upward bias in estimated genetic effects for the unconstrained approach. Likewise, there is an upward bias in estimated risk ratios for the constrained approach shown in Table 4. This bias was likely caused by violation of the assumption of generalized single ascertainment. With correlation among the parameters, the accuracy for location estimates was lowered by this bias in genetic effects. Thus, location estimates can be biased even with large sample sizes, depending on the ascertainment of pedigrees. For each model, we calculated the mean squared error, $MSE(\hat{\theta}) = (\hat{\theta} - \theta)^2 + SE^2(\hat{\theta})$, a composite measure of accuracy and precision for estimation. Since the genetic effects for our models are small to moderate, the corresponding transformed λ_{ik} parameters are similar across the types of ARPs. Given sufficient numbers of pairs for each type of ARP, we do not expect much difference in MSE between the unconstrained and the constrained model approaches. From Tables 3 and 4, by the MSE criterion, the unconstrained approach seems a bit better than the constrained approach for the four models. There are two reasons for this: First, when the recurrence risk ratios are close to 1 (small genetic effects), the iterations became even more laborious than that required for the unconstrained model. Once λ is estimated close to 1, genetic effects for all three kinds of ARPs reduce to 0, and there is no genetic information to be extracted. To avoid numerical problems caused by this, we set a lower bound 1.03 for λ estimates in each iteration. With similar risk ratios among ARPs (all close to 1), there is no severe violation if one resorts to the constrained approach. However, the lower bound for λ in the constrained approach might lead to slightly upward biased estimates. Goring et al. [12] also found that: "In general, estimates of bounded parameters are often biased, whether obtained by maximum likelihood or any other method." "In most cases, the closer to a boundary the true value of the parameter, the larger the bias from this source." This type of bias goes away asymptotically, yet, the accuracy for our location estimates might be affected by this upward bias in risk ratios. Second, we have sufficient numbers of pairs for each type of ARP, so the estimation of C_{ik} is stable, even if the unconstrained approach is used. To conclude, the unconstrained approach can be applied when there are sufficient numbers of pairs for each type of ARP. On the other hand, when data contains few pairs for some ARPs, the parsimonious constrained approach is better when the λ_{ik}'s are close to each other. The score test (7) used to evaluate the heterogeneity between λ_{ik}'s is valid in the sense that the type I error rate is as expected. We considered model E under no dominance and no epistasis, and mimicked a microsatellite scan for 100 families (details not shown). The empirical type I error rate is 4.98% for the nominal significance level of 5% for 1926 repetitions, showing that the score statistic (7) has a chi-square distribution in a moderate sample size.

4 Application to Prostate Cancer Linkage

A genome linkage scan of 157 families with multiple cases of prostate cancer was conducted by investigators at the Mayo Clinic by use of SNP and microsatellite markers in the Early Access Affymetrix Mapping 10K array [13]. The strongest linkage signal was detected on chromosome 20. Schaid et al. [2] analyzed the data on SNP markers by a one-locus localization method, and found it unclear whether there were two susceptibility loci on chromosome 20. We reanalyzed the data on 116 markers (13 microsatellite and 103 SNP markers) by one-locus and two-locus localization methods. The mean intermarker spacing was 0.97 cM, and the median was 0.5 cM. Among the 157 families, there were 279 full-sib pairs from 151 families, 115 first-cousin pairs from 38 families, and 26 avuncular pairs from 11 families. First we extracted each type of ARP and fit the two-locus model separately. The results were shown in Table 5 and Figure 3. We calculated the 95% CI for location parameter based on both the robust variance estimator and the bootstrap method, since the former usually underestimates the variation, suggested by our simulations. For full siblings and first cousins, we tried different sets of initial values and found consistent solutions for two loci and the corresponding genetic effects. Schaid et al. [2] found two different solutions in the FC subset, 24.1 and 92.3 cM, by the one-locus model. Our two-locus model found 23.6 and 98.9 cM simultaneously. This illustrates that omitting the second gene might bias the location estimate of the first gene. For avuncular pairs, we could not find the solutions that fit the two-locus model, so we resorted to the one-locus model. However, the location estimate was quite a distance from that obtained in the subsets of FS and FC. Since there were only

26 AP pairs, the result of the AP subset was inconclusive. Next we used all the data to fit the unconstrained and constrained models. In the unconstrained model, the AP pairs did not show a genetic effect at the common estimated τ_2, while the FC seemed to imply a second gene more distant from the first gene. However, the confidence intervals for the locations were quite wide and were overlapping for different types of ARPs. To evaluate the appropriateness of fitting a constrained model, we used the score test (7); the resultant test statistic was $T = 1.7$, with 4 df and a P-value of 0.79. This result suggested that there was no statistically significant difference among the λ_{1k}'s or λ_{2k}'s across the different types of ARPs, and thus a constrained model may be appropriate. The position estimates were similar in the unconstrained and constrained models. Compared with the single-locus results in Schaid et al. [2], their location estimates (72.9 for unconstrained and 72.7 for constrained model) might be shifted towards a second locus if there are actually two prostate-susceptibility genes located on chromosome 20.

5 Discussion

We have generalized the multipoint IBD mapping for two causative loci on a chromosomal region to allow for general kinds of ARPs. This can utilize pedigree data when they are available, and usually shortens the confidence intervals for gene locations by an increasing number of pairs. We have implemented these localization methods in our software GEEARP, for both one and two linked loci. To determine whether a two-locus model should be fit, a preliminary plot of average IBD scores along a chromosome for each type of ARP is useful. Note that when ARPs are analyzed separately or the unconstrained model is used, our asymptotic results will not be appropriate if the number of each type of ARP is not large. Although it is difficult to know the minimum number, we advise that at least 20 ARPs should be available for each type.

In summary, there are three sources of bias for this method. First, because this method finds positions and effects of genes that best fit the average IBD scores, it suffers from a downward bias for genetic effects when markers are not fully informative. GEEARP reduces this bias to some extent by first deleting completely noninformative pairs. However, because markers are not fully informative in most cases, this source of bias is unavoidable. Note that a general downward bias for genetic effects or risk ratios does not appear in Tables 3 and 4, because we generated fully informative markers in our simulations. Second, an upward bias of effect size can result from violation of the assumption of generalized single ascertainment, which can bias the estimation for locations as well. Because generalized single ascertainment is rarely used for linkage, particularly large pedigrees, our methods can suffer from this kind of bias, with the amount of bias depending on the true effect size and the ascertainment criteria. Third, an upward bias can occur when the true λ parameters are close to 1, i.e., small genetic effects, and a lower bound is placed on λ. To conclude, our simulations suggest that a model analyzed by the unconstrained approach suffers from the first two sources of bias, while the constrained approach suffers from all three sources of bias. With these complexities, it is hard to give a general rule for the direction of bias for genetic effects or risk ratios. Nonetheless, it appears that the downward bias from incomplete IBD information is more severe than the latter two.

Gene locations can be estimated by our methods, but their precision can be low because of the nature of coarse mapping in linkage analyses. Follow-up association studies in the confidence intervals are necessary for identifying causative loci. However, the CI coverage for gene location provided by the robust variance estimator usually does not attain the nominal size. Bootstrap corrects the downward bias of standard errors for location estimates, and its quantiles give more reliable confidence intervals. Because of large variability and hidden bias in IBD estimation, as well as modest genetic effects combined with small sample sizes, insufficient linkage information inflates the standard errors of parameter estimates and our method cannot have good performance. Without a large sample size and sufficient linkage evidence, it is hard to find convergent and consistent solutions, and it is inappropriate to resort to this complicated model.

Another caution in real data analyses is the bias that might occur because of linkage disequilibrium (LD), particularly when dense SNPs are analyzed. GEEARP accepts IBD output from Merlin that accounts for LD [14, 15]. A recent implementation of multipoint IBD mapping for rheumatoid arthritis-susceptibility genes on chromosome 6, both for one- and two-locus models, was analyzed by our accompanying software [16]. Comparing the two-locus model with the one-locus model [1, 2], if there are multiple loci in a region, the one-locus model performs well when the linkage peaks are far apart (close to the situation of unlinked genes) and there is no dominating peak. However, if there is a gene with much larger genetic effect than others, the one-locus model may fail to find other minor-linked loci. In contrast, the two-locus model allows us to find a second gene in the presence of a much stronger signal. However, the two-locus model can be over-parameterized and lose statistical efficiency if there is only one gene in a region. With these considerations, we suggest fitting both one- and two-locus models to evaluate linkage of multiple genes in a chromosomal region.

6 Acknowledgements

We are grateful for the helpful comments from the anonymous reviewers that improved this work. This research was supported by Taiwan's National Science Council, contract grant number NSC 095-2917-I-002-012 (WYL), and the U.S. Public Health Service, National Institutes of Health, contract grant number GM67768 (DJS).

7 Appendix

To solve Eq (2), because of using Haldane's mapping function, τ_1 and τ_2 are two change points. The quasi-likelihood is not differentiable when the change points are very close to marker positions. This leads to a tremendous step size in the iterations that causes the next evaluation to be outside the boundary, or even tend to infinity. We follow Liang et al. [1] by using a smooth curve when t and τ_1 or τ_2 are quite close. The following table summarizes our minor modification for computations.

Interval	$\boldsymbol{E}[S(t)\|\Phi]$	d_1	d_2	d_3
$t < \tau_1 - \epsilon$	$a_k + b_k(d_1)C_{1k}$	$\tau_1 - t$		
$\tau_1 - \epsilon < t < \tau_1 + \epsilon$	$a_k + b_k(d_1)C_{1k}$	$\frac{(t-\tau_1)^2}{2\epsilon} + \frac{\epsilon}{2}$		
$\tau_1 + \epsilon < t < \tau_2 - \epsilon$	$a_k + \frac{b_k(d_1)[1-b_k^2(d_2)]}{1-b_k^2(d_3)}C_{1k} + \frac{b_k(d_2)[1-b_k^2(d_1)]}{1-b_k^2(d_3)}C_{2k}$	$t - \tau_1$	$\tau_2 - t$	$\tau_2 - \tau_1$
$\tau_2 - \epsilon < t < \tau_2 + \epsilon$	$a_k + b_k(d_2)C_{2k}$		$\frac{(t-\tau_2)^2}{2\epsilon} + \frac{\epsilon}{2}$	
$t > \tau_2 + \epsilon$	$a_k + b_k(d_2)C_{2k}$		$t - \tau_2$	

However, this modification causes two gaps on the mean function. It seems implausible to find a curve that simultaneously smoothes the middle interval and the two sides. Despite the gaps on the mean function, the iteration process is improved by a small value of $\epsilon = 0.1$. A common situation is that the iterations run between two or several evaluations repeatedly. Decreasing the step size usually solves the difficulty. The two-locus model does not ensure convergence and consistent solutions, and one should try different sets of initial values to determine reliable solutions. The above numerical techniques have been implemented in our software, which is available by sending an email to WYL: d92842006@ntu.edu.tw

8 References

1 Liang K-Y, Chiu Y-F, Beaty TH: A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases. Hum Hered 2001;51:64-78.

2 Schaid DJ, Sinnwell JP, Thibodeau SN: Robust multipoint identical-by-descent mapping for affected relative pairs. Am J Hum Genet 2005;76:128-138.

3 Biernacka JM, Sun L, Bull SB: Simultaneous localization of two linked disease susceptibility genes. Genet Epidemiol 2005;28:33-47.

4 Risch N, Merikangas K: The future of genetic studies of complex human diseases. Science 1996;273:1516-1517.

5 Liang K-Y, Zeger SL: Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13-22.

6 Biernacka JM, Sun L, Bull SB: Tests for the presence of two linked disease susceptibility genes. Genet Epidemiol 2005;29:389-401.

7 Kong A, Cox NJ: Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 1997;61:1179-1188.

8 Haldane JBS: The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 1919;8:299-309.

9 Hodge SE, Vieland VJ: The essence of single ascertainment. Genetics 1996;144:1215-1223.

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

11 Risch N: Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 1990;46:222-228.

12 Goring HHH, Terwilliger JD, Blangero J: Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 2001;69:1357-1369.

13 Schaid DJ, Guenther JC, Christensen GB, Hebbring S, Rosenow C, Hilker CA, McDonnell SK, Cunningham JM, Slager SL, Blute ML, Thibodeau SN: Comparison of microsatellites versus single nucleotide polymorphisms by a genome linkage screen for prostate cancer susceptibility loci. Am J Hum Genet 2004;75:948-965.

14 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics 2002;30:97-101.

15 Abecasis GR, Wigginton JE: Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 2005;77:754-767.

16 Schaid DJ, Lin WY: One- and two-locus models for mapping rheumatoid arthritis-susceptibility genes on chromosome 6. In submission.

9 Tables

Table 1 Expected alleles shared IBD at location t for five types of ARPs and functions relating λ to C

Interval	$\boldsymbol{E}[S(t)\|\Phi]$	d_1	d_2	d_3
$t < \tau_1 < \tau_2$	$a_k + b_k(d_1)C_{1k}$	$\tau_1 - t$		
$\tau_1 < \tau_2 < t$	$a_k + b_k(d_2)C_{2k}$		$t - \tau_2$	
$\tau_1 < t < \tau_2$	$a_k + \frac{b_k(d_1)[1-b_k^2(d_2)]}{1-b_k^2(d_3)}C_{1k} + \frac{b_k(d_2)[1-b_k^2(d_1)]}{1-b_k^2(d_3)}C_{2k}$	$t - \tau_1$	$\tau_2 - t$	$\tau_2 - \tau_1$

ARP	a_k	$b_k(d)$	$C(\lambda)$	$\lambda(C)$
FS	1	$exp(-.04d)$	$\frac{\lambda-1}{2\lambda}$	$\frac{1}{1-2C}$
HS	$\frac{1}{2}$	$exp(-.04d)$	$\frac{\lambda-1}{2(\lambda+1)}$	$\frac{1+2C}{1-2C}$
FC	$\frac{1}{4}$	$\frac{1}{2}exp(-.04d) + \frac{1}{3}exp(-.06d) + \frac{1}{6}exp(-.08d)$	$\frac{3(\lambda-1)}{4(\lambda+3)}$	$\frac{12C+3}{3-4C}$
GP	$\frac{1}{2}$	$exp(-.02d)$	$\frac{\lambda-1}{2(\lambda+1)}$	$\frac{1+2C}{1-2C}$
AP	$\frac{1}{2}$	$\frac{1}{2}exp(-.04d) + \frac{1}{2}exp(-.06d)$	$\frac{\lambda-1}{2(\lambda+1)}$	$\frac{1+2C}{1-2C}$

This is the theoretical mean function for IBD scores, and we have a minor modification for computation described in the Appendix.

Table 2 The genetic models considered in simulation studies

Model	**A**			**B**			**C**			**D**		
Penetrance factors												
Locus 1	(0.02,0.25,0.40)			(0.05,0.05,0.45)			(0.20,0.70,1.00)			(0.10,0.50,0.90)		
Locus 2	(0.03,0.25,0.45)			(0.05,0.25,0.45)			(0.20,0.60,0.90)			(0.10,0.50,0.90)		
Penetrance matrix												
	aa	aA	AA	aa	aA	AA	aa	aA	AA	aa	aA	AA
bb	0.05	0.28	0.43	0.10	0.10	0.50	0.04	0.14	0.20	0.01	0.05	0.09
bB	0.27	0.50	0.65	0.30	0.30	0.70	0.12	0.42	0.60	0.05	0.25	0.45
BB	0.47	0.70	0.85	0.50	0.50	0.90	0.18	0.63	0.90	0.09	0.45	0.81
Allele frequencies												
	Pr(A)=0.05			Pr(A)=0.15			Pr(A)=0.10			Pr(A)=0.15		
	Pr(B)=0.05			Pr(B)=0.15			Pr(B)=0.10			Pr(B)=0.15		
Prevalence												
	9.48%			16.90%			8.31%			4.84%		
C_{ik} **coefficients when** $\tau_2 - \tau_1 = 40$ **cM**												
FS	0.106,0.101			0.040,0.078			0.110,0.094			0.177,0.177		
FC	0.051,0.048			0.007,0.032			0.046,0.038			0.083,0.083		
AP	0.061,0.058			0.010,0.041			0.058,0.048			0.099,0.099		
λ_{ik} **coefficients when** $\tau_2 - \tau_1 = 40$ **cM**												
FS	1.268,1.254			1.087,1.186			1.283,1.231			1.546,1.546		
FC	1.289,1.275			1.039,1.180			1.262,1.212			1.497,1.497		
AP	1.277,1.263			1.039,1.180			1.263,1.212			1.496,1.496		

1. A and B were two-locus additive models, while C and D were two-locus multiplicative models.
2. The two disease susceptibility genes were assumed to be diallelic with alleles (A,a) and (B,b), respectively.
3. The simulation contains three common types of ARPs: full siblings (FS), first cousins (FC), and avuncular pairs (AP).

Table 3. Parameter estimation for unconstrained-model approach

		Location				Genetic effects				95% CI	
		(cM)		FS		FC		AP		coverage (%)	
Model		τ_1	τ_2	C_{11}	C_{21}	C_{13}	C_{23}	C_{15}	C_{25}	τ_1	τ_2
A	Average estimate	33.7	75.3	0.117	0.115	0.048	0.047	0.083	0.084	69	68
	Average robust SE	4.5	4.4	0.041	0.040	0.049	0.048	0.057	0.056		
	Empirical SE	8.3	8.2	0.040	0.038	0.053	0.051	0.059	0.060		
	Mean Bias	-1.3	0.3	0.011	0.014	-0.003	-0.001	0.022	0.026		
	MSE	70.6	67.3	0.002	0.002	0.003	0.003	0.004	0.004		
	Bootstrap estimate	35.7	75.9	0.119	0.125	0.056	0.048	0.074	0.069	96	100
	Bootstrap SE	9.5	8.9	0.041	0.042	0.054	0.055	0.059	0.058		
B	Average estimate	36.2	76.3	0.055	0.091	0.011	0.031	0.019	0.058	59	72
	Average robust SE	6.5	5.0	0.041	0.040	0.045	0.045	0.051	0.050		
	Empirical SE	14.2	8.2	0.045	0.038	0.054	0.048	0.061	0.051		
	Mean Bias	1.2	1.3	0.015	0.013	0.004	-0.001	0.009	0.017		
	MSE	203.1	68.9	0.002	0.002	0.003	0.002	0.004	0.003		
	Bootstrap estimate	33.8	74.7	0.051	0.083	0.007	0.028	0.009	0.043	98	98
	Bootstrap SE	12.0	9.1	0.043	0.041	0.052	0.049	0.058	0.053		
C	Average estimate	33.6	74.3	0.125	0.111	0.046	0.042	0.088	0.075	69	69
	Average robust SE	4.4	4.8	0.041	0.041	0.048	0.048	0.060	0.060		
	Empirical SE	7.7	8.6	0.039	0.041	0.050	0.052	0.061	0.063		
	Mean Bias	-1.4	-0.7	0.015	0.017	0.000	0.004	0.030	0.027		
	MSE	61.3	74.5	0.002	0.002	0.003	0.003	0.005	0.005		
	Bootstrap estimate	35.1	74.7	0.114	0.094	0.037	0.041	0.067	0.058	93	95
	Bootstrap SE	8.5	8.6	0.039	0.042	0.053	0.052	0.058	0.060		
D	Average estimate	34.0	75.4	0.184	0.180	0.083	0.078	0.135	0.128	73	73
	Average robust SE	3.1	3.1	0.040	0.040	0.054	0.053	0.066	0.065		
	Empirical SE	5.7	5.4	0.038	0.037	0.055	0.055	0.064	0.066		
	Mean Bias	-1.0	0.4	0.007	0.003	0.000	-0.005	0.036	0.029		
	MSE	33.5	29.3	0.001	0.001	0.003	0.003	0.005	0.005		
	Bootstrap estimate	33.7	74.1	0.187	0.179	0.078	0.081	0.098	0.105	97	93
	Bootstrap SE	8.6	6.7	0.046	0.045	0.050	0.051	0.063	0.063		

Table 4. Parameter estimation for constrained-model approach

		Location (cM)				95% CI coverage (%)	
Model		τ_1	τ_2	λ_1	λ_2	τ_1	τ_2
A	Average estimate	33.4	74.7	1.324	1.311	74	76
	Average robust SE	5.4	5.5	0.128	0.125		
	Empirical SE	9.5	8.9	0.102	0.098		
	Mean Bias	-1.6	-0.3	0.056	0.057		
	MSE	92.8	79.3	0.014	0.013		
	Bootstrap estimate	35.9	77.3	1.321	1.287	91	97
	Bootstrap SE	10.3	9.9	0.105	0.107		
B	Average estimate	37.3	77.3	1.197	1.244	73	78
	Average robust SE	9.1	6.1	0.105	0.110		
	Empirical SE	13.4	8.4	0.070	0.075		
	Mean Bias	2.3	2.3	0.110	0.058		
	MSE	184.9	75.9	0.017	0.009		
	Bootstrap estimate	33.7	73.3	1.248	1.266	90	100
	Bootstrap SE	12.2	10.0	0.088	0.081		
C	Average estimate	33.2	74.0	1.330	1.304	72	72
	Average robust SE	5.4	5.8	0.132	0.127		
	Empirical SE	9.6	10.0	0.102	0.100		
	Mean Bias	-1.8	-1.0	0.047	0.073		
	MSE	95.4	101.0	0.013	0.015		
	Bootstrap estimate	33.4	73.2	1.314	1.356	93	91
	Bootstrap SE	10.7	9.1	0.128	0.133		
D	Average estimate	35.3	71.7	1.633	1.641	60	58
	Average robust SE	3.2	3.0	0.190	0.191		
	Empirical SE	7.3	6.3	0.176	0.180		
	Mean Bias	0.3	-3.3	0.087	0.095		
	MSE	53.4	50.6	0.039	0.041		
	Bootstrap estimate	34.2	74.5	1.504	1.519	96	96
	Bootstrap SE	9.3	7.3	0.142	0.152		

1. True values for $\tau_1 = 35$ and $\tau_2 = 75$ (cM). The initial values in the Fisher's scoring method were set at $\tau_1 = 50, \tau_2 = 60, C's = 0.1$ in the unconstrained model and $\lambda's = 1.2$ in the constrained model.

2. The average number of (ARPs = FS + FC + AP) is (669 = 424 + 130 + 115).

3. The 95% CIs based on robust variance estimator were calculated by estimate±1.96×(robust SE).

4. MSE($\hat{\tau}_1$)=$(\hat{\tau}_1 - \tau_1)^2$+$SE^2(\hat{\tau}_1)$. Mean bias and MSE were based on 1,000 non-bootstrap repetitions.

5. Bootstrap estimates were based on 100 repetitions, in each repetition, we had 500 bootstrap samples drawn with replacement from 500 families ($n = B = 500$). Bootstrap SE was the mean of

100 empirical standard errors calculated from the 100 repetitions. The 95% CIs based on bootstrap method were calculated by (2.5-97.5) percentiles of bootstrap estimates.

6. Strictly speaking, no "bias" could be calculated for risk ratios in Table 4, because none of the models in Table 2 was under no dominance and no epistasis. Here we use $\hat{\lambda_i} - \lambda_{i1}$ as a surrogate.

Table 5. Parameter Estimates for Prostate Cancer Linkage

TYPE OF ANALYSIS	ESTIMATE (95% CI) FOR			
AND ARP	τ_1	τ_2	λ_1^a	λ_2
Subsets^b**:**				
Full siblings	26.5 (13.8-39.2; 4.4-53.4)^c	73.1 (63.5-82.7; 59.5-90.5)	1.2 (0.9-1.4)	1.4 (1.0-1.8)
First cousins	23.6 (11.6-35.5; 1.5-52.5)	98.9 (90.9-106.9; 55.5-111.5)	1.5 (0.6-2.3)	1.8 (1.1-2.5)
Avuncular pairs	12.4 (0.0-43.7; 0.0-55.3)		1.7 (0.0-4.4)	
Unconstrained Model^d**:**				
Full siblings			1.2 (1.0-1.4)	1.4 (1.0-1.8)
First cousins			1.5 (0.6-2.4)	1.4 (0.7-2.1)
Avuncular pairs			1.5 (0.0-3.3)	0.9(0.0-1.8)
Common τ_1, τ_2	24.3 (17.4-31.1; 8.8-44.8)	75.9 (66.5-85.3; 53.7-98.3)		
Constrained Model^e**:**	25.9 (17.3-34.5; 12.4-41.8)	76.1 (66.3-85.9; 63.8-92.0)	1.2 (1.0-1.5)	1.4 (1.0-1.7)

[a] All the C coefficients were transformed into λ for comparisons.

[b] Subsets were fit for different types of ARPs separately. For avuncular pairs, there was only one locus to be found.

[c] The former CI was based on the robust variance estimates, while the latter was based on (2.5-97.5) percentiles of 1,000 bootstrap estimates.

[d] The unconstrained model was fit by combining all types of ARPs, estimating common τ_1 and τ_2 with different genetic effects for the three types of ARPs.

[e] The constrained model was fit by combining all types of ARPs, estimating common τ_1, τ_2, λ_1, and λ_2.

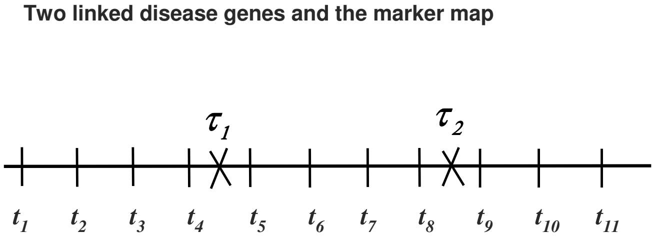

Figure 1. Hypothetical locations of 11 observed markers $(t_1, t_2, \cdots, t_{11})$ and two unobserved linked disease susceptibility genes (τ_1, τ_2) in a chromosomal region. Each marker was 10 cM apart from the adjacent one. The two causative loci were located at 35 cM and 75 cM, respectively.

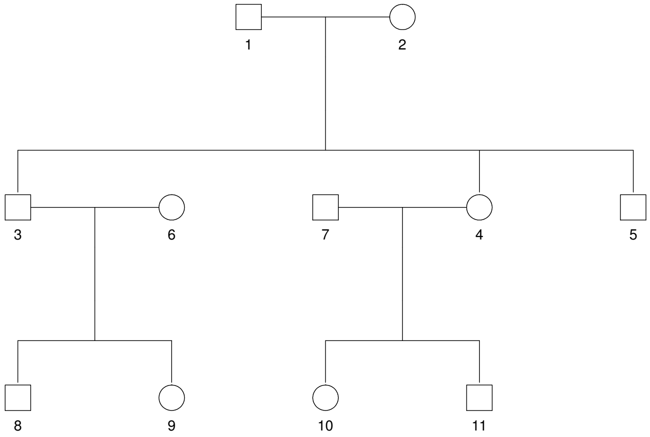

Figure 2. Structure of pedigrees used for simulation study. Only three types of ARP: full siblings, first cousins, and avuncular pairs, were included in simulation.

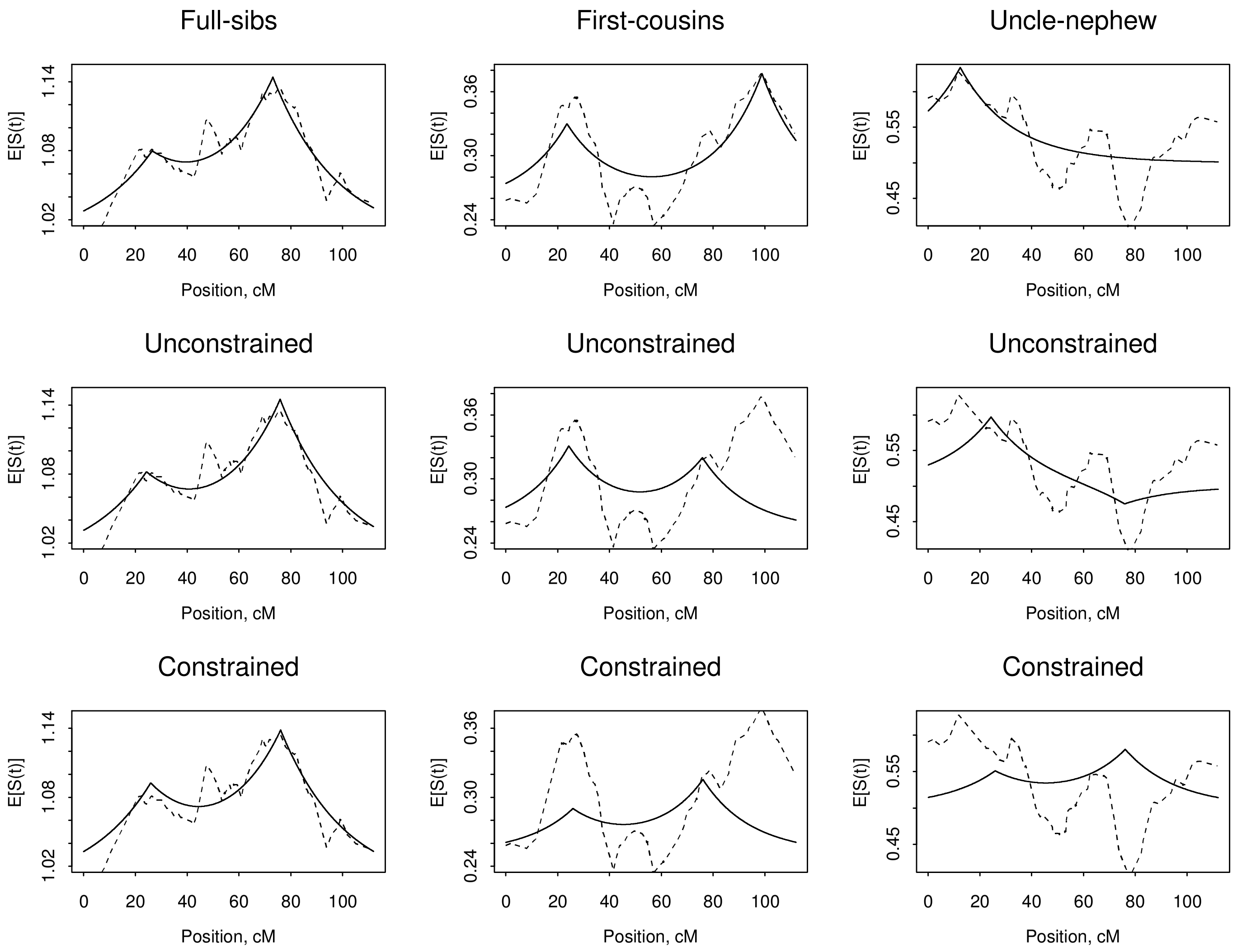

Figure 3. Average IBD scores (*broken line*) and fitted values (*solid line*) for FS, FC, and AP, respectively (*columns*). The top row is for subset analysis, the second row is for unconstrained model, while the bottom row is for constrained model results.