Simultaneous Localization of Two Linked Disease Susceptibility Genes

Joanna M. Biernacka,1,2 Lei Sun,1,3 and Shelley B. Bull1,2*

1Department of Public Health Sciences, University of Toronto, Toronto, Canada
2Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada
3The Hospital for Sick Children, Toronto, Canada

For diseases with complex genetic etiology, more than one susceptibility gene may exist in a single chromosomal region. Extending the work of Liang et al. (2001) Hum. Hered. 51:64–78, we developed a method for simultaneous localization of two susceptibility genes in one region. We derived an expression for expected allele sharing of an affected sib pair (ASP) at each point across a chromosomal segment containing two susceptibility genes. Using generalized estimating equations (GEE), we developed an algorithm that uses marker identical-by-descent (IBD) sharing in affected sib pairs to simultaneously estimate the locations of the two genes and the mean IBD sharing in ASPs at these two disease loci. Confidence intervals for gene locations can be constructed based on large sample approximations. Application of the described methods to data from a genome scan for type 1 diabetes (Mein et al. 1998) Nat. Genet. 19:297–300) yielded estimates of two putative disease gene locations on chromosome 6, approximately 20 cM apart. Properties of the estimators, including bias, precision, and confidence interval coverage, were studied by simulation for a range of genetic models. The simulations demonstrated that the proposed method can improve disease gene localization and aid in resolving large peaks when two disease genes are present in one chromosomal region. Joint localization of two disease genes improves with increased excess allele sharing at the disease gene loci, increased distance between the disease genes, and increased number of affected sib pairs in the sample. Genet. Epidemiol. 28:33–47, 2005. © 2004 Wiley-Liss, Inc.

Key words: generalized estimating equations; affected sib pairs; identity by descent; complex disease; multi-locus

INTRODUCTION

Various human traits and rare diseases, controlled by single genes and having simple Mendelian inheritance, have been successfully studied for some time. Recently, however, research in human genetic disease has focused on investigating the more complex, but also more prevalent oligogenic or multifactorial traits, which may be controlled by multiple genetic and environmental factors, as well as their interactions. This has led to increased interest in methods that incorporate the fact that there may be more than one gene contributing to the disease.

Methods that take into account the role of multiple genes and environmental factors may be more powerful and improve the precision of disease-gene localization. When developing and assessing such methods, it is important to identify the main purpose of the analysis: detection of linkage, estimation of the locations of disease genes, or characterization of the type of interaction between genes. In order to study the interactions between genes, however, one must first identify those genes and find their positions in the genome. Thus, precise localization of disease susceptibility genes is of primary importance.

A common feature of most of the two-locus linkage methods described in the literature is that they are designed to deal only with unlinked disease genes. The problem of two linked loci has not been examined extensively in the literature, excepting the work of Delépine [1997], Farrall [1997], Cordell et al. [1998], and Biswas et al.

© 2004 Wiley-Liss, Inc.
Cordell et al. [1998] described association methods for mapping multiple linked QTLs in inbred animal populations. The focus of Farrall’s [1997] work was to resolve genetic interactions between linked loci and improve power to detect linkage to a second gene, by taking into account evidence for linkage at a linked gene with an established location. This general approach was applied to IDDM data by Cordell et al. [2000], showing some evidence for a second type 1 diabetes gene on chromosome 6 linked to IDDM1. Delépine et al. [1997] also developed a method to test for the presence of a second susceptibility gene linked to a known susceptibility gene and found evidence for a second type 1 diabetes susceptibility gene on chromosome 6 linked to HLA. When the two loci are linked, localization of the first locus could be inaccurate because it is mapped without taking into account the presence of linked trait genes. Then, conditioning on an inaccurate position for the first gene may lead to poor estimates of the second gene location. The alternative approach of simultaneously localizing two linked disease genes has the potential to improve the location estimates of both genes. Biswas et al. [2003] applied a Bayesian approach to the simultaneous detection of two linked disease genes. This method, however, is a parametric approach requiring the specification of the penetrance vectors, designed to detect genes under locus heterogeneity.

In the absence of good candidate genes, multipoint linkage analysis is the most commonly used technique for finding disease susceptibility genes. Allele sharing methods have the advantage that they do not require the assumption of a specific genetic model with, for example, recessive or dominant mode of inheritance. However, most of these methods are not designed to estimate the locations of disease genes with a specified level of certainty, since they do not provide confidence intervals for disease gene locations. Recently, Liang et al. [2001] introduced a generalized estimating equations (GEE) method to estimate the location of a trait gene along with a confidence interval for its map position. Under the assumptions of random mating, linkage equilibrium, and generalized single ascertainment, they showed that in a chromosomal region with one susceptibility gene at location \( t \):

\[
\mu(t) = E(S(t)|\Phi) = 1 + (2\Psi_{1,t} - 1)C
\]

where \( C = E(S(t)|\Phi) - 1 \). \( S(t) \) = number of alleles (0, 1, or 2) shared IBD by an ASP at locus \( t \), \( \Phi \) denotes the event that both sibs are affected, and \( \Psi_{1,t} = \theta_{1,t}^2 + (1 - \theta_{1,t})^2 \), where \( \theta_{1,t} \) is the

METHODS

Liang et al. [2001] proposed an IBD-based procedure to estimate the location of a susceptibility gene within a chromosomal region framed by multiple markers. Under the assumptions of random mating, linkage equilibrium, and generalized single ascertainment, they showed that in a chromosomal region with one susceptibility gene at location \( t \):
Localization of Linked Disease Genes

recombination fraction between \( t \) and \( \tau \). Liang et al. [2001] proposed using marker IBD sharing for a sample of ASPs to estimate the location of the disease gene, \( \tau \), and \( C = E(S(\tau) | \Phi) - 1 \) by a generalized estimating equations (GEE) approach, and implemented an algorithm to estimate the parameters \( C \) and \( \tau \) in a program called GENE-FINDER.

A MODEL FOR MEAN ALLELE SHARING IN A REGION WITH TWO DISEASE LOCI

Let \( \tau_1 \) and \( \tau_2 \) be the locations of two linked trait loci \( (\tau_1 < \tau_2) \). Again, let \( S(t) \) be the number of alleles \( (0, 1, \) or \( 2) \) shared IBD by an ASP at locus \( t \), and \( \Phi \) denote the event that both sibs are affected. Define \( C_1 = E(S(\tau_1) | \Phi) - 1 \) and \( C_2 = E(S(\tau_2) | \Phi) - 1 \). Also let \( \Psi_1 = \Psi_{t,\tau_1}, \Psi_2 = \Psi_{t,\tau_2} \) and \( \Psi_3 = \Psi_{t_1,\tau_2} (\Psi_{A,B} = \theta_{A,B,t}^2 + (1-\theta_{A,B})^2) \), where \( \theta_{A,B,t} \) is the recombination fraction between loci \( A \) and \( B \). Under the assumptions of random mating, linkage equilibrium, generalized single ascertainment, no interference, and equality of recombination fractions for males and females, \( E(S(t) | \Phi) \) can be written as a function of \( C_1, C_2, \Psi_1, \Psi_2, \) and \( \Psi_3 \) (proof in the Appendix).

Using Haldane’s mapping function to translate recombination fraction to map distance \( \theta_{1,1} = (1 - e^{-0.02|t-t_1|})/2 \), \( \mu_2(t) = E(S(t) | \Phi) \) can be rewritten in terms of \( t \) and \( \delta_2=(C_1, C_2, \tau_1, \tau_2) \). We use \( \mu_2(t; \delta_2) \) to denote this two-locus model mean function for all \( t \).

For \( t < \tau_1 < \tau_2 \): \( \mu_2(t; \delta_2) = E(S(t) | \Phi) = 1 + (2\Psi_1 - 1) \cdot C_1 + C_2 e^{-0.04|t-\tau_1|} \)

for \( \tau_1 < t < \tau_2 \): \( \mu_2(t; \delta_2) = E(S(t) | \Phi) = 1 + (2\Psi_2 - 1) \cdot C_2 + C_1 e^{-0.04|t-\tau_1|} \)

and for \( \tau_1 < t < \tau_2 \):

\[
\mu_2(t; \delta_2) = E(S(t) | \Phi) = 1 + \frac{(2\Psi_1 - 1)(1 - (2\Psi_2 - 1)^2)}{1 - (2\Psi_3 - 1)^2} C_1 \]

\[
+ \frac{(2\Psi_2 - 1)(1 - (2\Psi_1 - 1)^2)}{1 - (2\Psi_3 - 1)^2} C_2 \]

\[
= 1 + C_1 e^{-0.04(t-\tau_1)} \frac{1 - e^{-0.08(t-\tau_1)}}{1 - e^{-0.08(t-\tau_1)}} \cdot C_2 e^{-0.04(t-\tau_1)} \cdot \frac{1 - e^{-0.08(t-\tau_1)}}{1 - e^{-0.08(t-\tau_1)}} \cdot \]

As in the single-locus model, the above formula for \( E(S(t) | \Phi) \) is robust in the sense that it holds regardless of the mode of inheritance. The parameters \( C_1 \) and \( C_2 \) depend on the underlying genetic mechanism, and \( E(S(t) | \Phi) \) depends on the genetic model only through the parameters \( C_1 \) and \( C_2 \) and the relative positions of \( t \), \( \tau_1 \), and \( \tau_2 \). No assumptions are made about the existence of other unlinked trait genes. Figure 1 shows the \( E(S(t) | \Phi) \) curve for a chromosomal segment containing two disease genes under two different genetic models. It is apparent that localization of a disease gene with lower IBD sharing located near a gene with high IBD sharing is likely to be more difficult than localization of two genes with similar levels of IBD sharing.

Note that when \( t \) is not linked to either of the two disease genes, then \( \Psi_1 = \Psi_{t,\tau_1} = 0.5, \Psi_2 = \Psi_{t,\tau_2} = 0.5 \) and \( E(S(t) | \Phi) = 1 \) as expected under the null hypothesis of no linkage to any disease genes. Also, if \( \tau_1 \) and \( \tau_2 \) are not linked, and \( t \) is linked only to \( \tau_1 \), then \( \Psi_2 = \Psi_{t,\tau_2} = 0.5 \) and \( E(S(t) | \Phi) = 1 + (2\Psi_1 - 1) \cdot C_1 \) for all \( t \). That is, the two-locus model reduces to the one-locus model in this special case.

Fig. 1. Plot of \( E[S(t) | \Phi] \) for (a) model A from Table I with \( \tau_1 = 30, \tau_2 = 60 \) \( (C_1 = C_2 = 0.161) \) and (b) model D from Table I with \( \tau_1 = 30, \tau_2 = 60 \) \( (C_1 = 0.348, C_2 = 0.214) \).
ESTIMATION OF MODEL PARAMETERS

In the two-locus model, \( \mu_2(t; \delta_2) \) is characterized parametrically by four parameters: \( C_1, C_2, \tau_1, \) and \( \tau_2 \). If a sample of \( n \) independent ASPs were collected, and IBD sharing for all ASPs at all marker loci \( \{t_1, \ldots, t_M \} \) were known, one could obtain estimates of the parameters \( \delta_2 = (C_1, C_2, \tau_1, \tau_2) \) by solving the GEE

\[
\sum_{i=1}^{n} \left( \frac{\partial \mu_2}{\partial \delta_2} \right) \right)^T \text{Cov}^{-1}(S_i|\Phi) (S_i - \mu_2(\delta_2)) = 0 \quad (1)
\]

where \( S_i = (S_i(t_1), \ldots, S_i(t_M))^T = (S_{i1}, \ldots, S_{iM})^T \) and \( \mu_2(\delta_2) = (\mu_2(t_1; \delta_2), \ldots, \mu_2(t_M; \delta_2))^T \). If there are two disease genes in the region, solving these equations for \( \delta_2 \) should provide consistent estimates of the parameters and the variances of the estimates because \( E[S_i] = \mu_2(\delta_2) \) for ASPs under the two-locus model.

However, in most cases, exact IBD sharing is not known for all ASPs at all markers. Assume an ASP design with \( M \) markers at positions \( t_1, \ldots, t_M \). For each sib-pair \( Y_i = (Y_i(t_1), \ldots, Y_i(t_M)) \) is observed, where \( Y_i(t_i) \) represents the marker information at locus \( t_i \) \( (j=1, \ldots, M) \) for the \( i \)th ASP family. If the \( j \)th marker is not fully informative for family \( i \), \( S_i(t_i) \) is unknown. In such cases, allele sharing can be imputed using all marker data for the \( i \)th family: \( S_i^*(t_i) = E[S_i(t_i)|Y_i], j = 1, \ldots, M \), for example, using the software program GENEHUNTER [Kruglyak et al., 1996]. Following the methods of Liang et al. [2001] for the single-locus-model, estimates of the two-locus-model parameters \( \delta_2 = (C_1, C_2, \tau_1, \tau_2) \) can be obtained by substituting the estimated IBD allele sharing \( S_i^* = (S_{i1}^*, S_{i2}^*, \ldots, S_{iM}^*) \) where \( S_{im}^* = S_{im}^*(t_m) \) for \( S_i \) in the GEE (equation 1) shown above and solving the resulting equations. The parameter estimates and their estimated precision will be consistent provided \( E[S_i^*|\Phi] = E[S_i|\Phi] = \mu_2(\delta_2) \). Assuming fully informative markers, we have \( E[S_i^*|\Phi] = \mu_2(\delta_2) \) since at any fully informative marker \( m \), \( S_{im}^* = S_{im} \). At non-fully-informative markers linked to one or more disease genes, \( S_{im}^* \) will tend to underestimate \( S_{im} \) because \( S_{im}^* = E[S_{im}|Y_i, H_0] \) is determined under the null hypothesis of no linkage to any disease genes. That is, \( S_{im}^* \) is the expected sharing at marker \( m \), conditional on all the marker data, computed without taking the phenotype of the sib-pairs into account. However, when the markers are highly (if not fully) informative, the expression will still hold approximately: \( E[S_i^*|\Phi] \approx \mu_2(\delta_2) \).

Estimates of the disease gene locations, \( \tau_1 \) and \( \tau_2 \), are of primary interest. Values of \( C_1 \) and \( C_2 \) depend on the underlying genetic model (penetrance matrix and allele frequencies at the two disease genes) and the distance between the disease genes. Although estimates of \( C_1 \) and \( C_2 \) are of limited use in the sense that knowledge of their values does not identify the underlying genetic model, their magnitude influences the accuracy and precision with which we can estimate the disease gene locations.

Although misspecification of \( \text{Cov}(S_i^*|\Phi) \) can reduce the asymptotic efficiency of the parameter estimates, under mild regularity conditions, a solution to the generalized estimating equations provides consistent estimates of the parameters \( \delta_2 \) even when the covariance matrix for the data is incorrectly specified [Liang and Zeger, 1986]. Furthermore, although downward bias of standard errors obtained from the robust variance estimator has been observed in small samples [e.g., Pan 2001], in large samples the robust sandwich estimator [White, 1982; Liang and Zeger, 1986] gives consistent estimates of the variances of parameter estimates. In this implementation, we assume an independence correlation structure and empirically estimate the diagonal elements of the covariance matrix, \( \text{Var}(S_i^*(t_i)|\Phi) \) (see the Appendix for details).

By large-sample quasi-likelihood theory, solving the generalized estimating equations results in consistent and asymptotically normally distributed parameter estimates. Thus, asymptotic 100(1-\( \alpha \))% confidence intervals (CIs) for the location estimates can be constructed as: \( \text{estimate} \pm z_{\alpha/2} \cdot \text{se} \), where \( z_{\alpha/2} \) is the 100(1-\( \alpha \))th percentile of the standard normal distribution, and \( \text{se} \) is the robust estimate of the standard error of the parameter estimate.

We implemented a solution to the two-locus model GEE in a FORTRAN program with a modified Fisher’s scoring algorithm. We applied the proposed method for localization of two linked trait loci to chromosome 6 data from a genome scan for type 1 diabetes [Mein et al., 1998], and performed a simulation study to evaluate properties of the estimates, including bias, efficiency, and confidence interval coverage.

APPLICATION TO DIABETES DATA

Mein et al. [1998] presented results of a genome scan for type 1 diabetes, showing very strong
evidence of linkage in a region containing the well-established IDDM1 gene(s) in the MHC/HLA region. Using the same data, Cordell et al. [2000] found evidence for a second type 1 diabetes gene linked to the HLA region. We fit the one- and the two-locus models to this data set. Figure 2 shows the observed marker IBD sharing of the 356 ASPs from the sample, along with the curves defined by estimates from the one-locus and two-locus models. Assuming there is only one disease gene in the region, the location of the single putative disease gene was estimated to be 28.0 cM, with an estimated standard error of 0.74. Expected IBD sharing (C+1) was estimated to be 1.47 (s.e. 0.04). With the two-locus model, putative disease gene locations were estimated to be 27.0 cM (s.e. 0.52) and 48.6 cM (s.e. 5.64). Expected IBD sharing in affected sib-pairs was estimated to be 1.45 (s.e. 0.03) and 1.26 (s.e. 0.04) at these two loci, respectively. Note that the fact that the 95% confidence interval for C2 (0.18–0.34) excludes zero does not imply significant evidence for at least two loci. Even in the absence of a second disease gene in the vicinity of the location τ2, excess IBD sharing is expected at τ2, as a result of the effect of the disease gene at τ1 on IBD sharing in this region. The one-locus-model asymptotic-theory 95% confidence interval for the disease gene location is (26.6, 29.5). The 95% CIs for the disease gene locations under the two-locus model are (26.0, 28.0) and (37.4, 59.5). The confidence interval for the second gene location excludes the IDDM1 locus, but is wider because the effect size of the second putative gene is low relative to the first gene.

Previously, Delépine et al. [1997] and Cordell et al. [2000] found evidence of a second disease gene for type 1 diabetes linked to the HLA region. However, the positions reported in these two studies were about 20 cM apart. Using the same data as Cordell et al., we obtained a confidence interval for the second disease gene location, which includes the location for the second gene reported by Cordell et al. [2000], but not the second gene reported by Delépine et al. [1997]. Our confidence interval was calculated based on asymptotic theory, and may not have nominal coverage in moderate-sized samples. Therefore, we carried out a simulation study to evaluate the performance of our estimation method, including bias of point estimates and confidence interval coverage.

**MONTE CARLO SIMULATION STUDIES**

**METHODS**

Simulations were carried out to assess properties of the proposed estimation method including bias and precision of parameter estimates, accuracy of standard error estimates, and confidence interval coverage. Simulation settings were chosen to study the effects of sample size, map properties such as number of markers and intermarker distances, and various aspects of the underlying genetic model including the magnitude of the C parameters, and distance between the two disease genes. Here, we present the results of four sets of simulations. Each simulation is based on 1,000 replicates. For the first three sets of simulations, we generated fully informative marker IBD sharing for ASPs under a variety of two-locus models using our own data generation program, while in a fourth set we considered less-than-fully-informative ASP data generated using Allegro [Gudbjartsson et al., 2000]. The penetrance matrix, disease allele frequencies, distances between the two disease genes, and the corresponding values of the excess sharing parameters C1 and C2 for each model used for generating fully informative data are shown in Table I. The models were chosen to represent a range of genetic models including heterogeneity, epistasis, and additivity.
SIMULATION STUDY I: DIFFERENT GENETIC MODELS

In the first set of simulations, we studied bias and standard deviations of parameter estimates, magnitude of their standard errors, and confidence interval coverage for four different underlying genetic models (Table I). In addition, using the same four models, we investigated the consequences of model misspecification, i.e., the results obtained by assuming only one disease gene in the region when there are two linked disease genes. For this set of simulations, IBD sharing for 500 independent ASPs was generated at 11 markers spaced at 10-cM intervals (i.e., at 0, 10, ..., 90, 100 cM), assuming the two disease genes are 30 cM apart.

Averages of the four parameter estimates in these simulations are very close to the true values (Table II), indicating little or no bias. Comparison of the average robust standard error estimates, obtained from the sandwich estimator, to the empirical standard deviations computed across simulation replicates (Table II) demonstrates that the robust variance estimates for the estimates of $\tau_1$ and $\tau_2$ tend to be too low. Therefore, using these variance estimates to construct confidence intervals (assuming an asymptotic normal distribution of the parameters), produces confidence intervals having lower than the nominal confidence levels. However, confidence interval coverage reaches the nominal level for models with larger underlying $C$ values (i.e., greater excess sharing at the two disease genes). Not surprisingly, performance is better (i.e., lower bias of estimates, smaller variances of estimates, CI coverage closer to nominal) when there is greater excess allele sharing in affected sib pairs at the two disease genes.

When we fit a one-disease-locus model to data generated under a model with two linked disease loci, on average, the estimate of $\tau$ falls between the two true disease gene locations and the robust variance estimate is substantially biased down, resulting in confidence intervals with low coverage of the two true locations (Table III). We also observed that when an incorrect model is fit to the data, solutions to the estimating equations are more dependent on the initial parameter values provided for the estimation algorithm. In four of the simulations presented in Table III, the initial value for $\tau$ was set to be the position mid-way between the two true locations (50 cM), which led to estimates near that initial value and confidence intervals for $\tau$ with very low coverage of the two true disease gene locations. When the initial value for $\tau$ was set to be one of the two true locations, estimates of $\tau$ tended to be closer to the initial value for $\tau$, and confidence intervals were more likely to include that disease locus (rows 2 and 5 of Table III). However, coverage of that disease gene locus was still low, and large bias remained in the estimate of its effect size, $C$.

---

TABLE I. Models used for data generation for simulations presented in Tables II–V

<table>
<thead>
<tr>
<th>Model</th>
<th>Penetrance matrix</th>
<th>Allele frequencies</th>
<th>Prevalence</th>
<th>$\tau_2-\tau_1$</th>
<th>$C_1$, $C_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$aa$ $aA$ $AA$</td>
<td>$bb$ 0.050 0.050 0.950</td>
<td>Pr(A)=.05</td>
<td>Approx 5%</td>
<td>30 0.161, 0.161</td>
</tr>
<tr>
<td></td>
<td>$bB$ 0.050 0.050 0.950</td>
<td>Pr(B)=.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$BB$ 0.950 0.950 0.950</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>$aa$ $aA$ $AA$</td>
<td>$bb$ 0.010 0.010 0.800</td>
<td>Pr(A)=.015</td>
<td>Approx 1%</td>
<td>10 0.327, 0.327</td>
</tr>
<tr>
<td></td>
<td>$bB$ 0.010 0.010 0.800</td>
<td>Pr(B)=.015</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$BB$ 0.800 0.800 0.800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>$aa$ $aA$ $AA$</td>
<td>$bb$ 0.000 0.000 0.000</td>
<td>Pr(A)=.053</td>
<td>Approx 1%</td>
<td>30 0.487, 0.487</td>
</tr>
<tr>
<td></td>
<td>$bB$ 0.000 0.950 0.950</td>
<td>Pr(B)=.053</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$BB$ 0.000 0.950 0.950</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>$aa$ $aA$ $AA$</td>
<td>$bb$ 0.005 0.005 0.400</td>
<td>Pr(A)=.02</td>
<td>Approx 0.5%</td>
<td>30 0.348, 0.214</td>
</tr>
<tr>
<td></td>
<td>$bB$ 0.005 0.005 0.400</td>
<td>Pr(B)=.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$BB$ 0.250 0.250 0.600</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SIMULATION STUDY II: SAMPLE SIZE AND DISTANCE BETWEEN LOCI

The second set of simulations was designed to demonstrate the effect of sample size and distance between the two disease genes on the estimates in a two-locus model. Disease genes were placed 20, 30, or 50 cM apart. IBD sharing data at 11 markers spaced at 10-cM intervals were generated for samples of 200, 500, and 1,000 independent ASPs.

With a larger sample size, we found less bias in estimates of the disease gene locations, and the variances of the estimates of $t_1$ and $t_2$ were lower, leading to narrower confidence intervals with closer to nominal coverage (Table IV). Greater bias in location estimates and less than nominal confidence interval coverage were observed when the two disease genes were closer together.

SIMULATION STUDY III: DENSITY AND NUMBER OF MARKERS

Marker densities and total map distances were varied in the third set of simulations (Table V). For each replicate, data were generated for 500 independent ASPs under the same penetrance and allele frequency model. Comparison of the first three rows of Table V indicates that while genotyping additional markers between the two disease loci may improve localization, having too many markers in the analysis can be detrimental in terms of confidence interval coverage. Further comparisons with other results presented in Table V establish that it is important to have several markers (3 or 4) between the two disease genes. However, genotyping additional markers far outside the region of the two disease genes may not be helpful and can also decrease confidence interval coverage (see discussion below).

SIMULATION STUDY IV: MARKER INFORMATIVENESS

The fourth set of simulations evaluated data sets with different levels of marker informativeness. For these simulations, genotype data were generated using Allegro [Gudbjartsson et al., 2000] assuming a locus heterogeneity model. For each replicate, data were generated for 250 ASPs and their parents from each of two populations, for a total sample of 500 ASPs. In the first population, there was a single disease gene at position $t_1=35$ cM with expected IBD sharing of 1.25 in ASPs at this disease gene. In the second population, there was a single disease gene at position $t_2=65$ cM.
also with expected IBD sharing of 1.325. Thus, in a population of ASPs consisting of equal numbers from each of the two populations, there are disease susceptibility loci at 35 and 65 cM, with $C_1 = C_2 = 0.211$. Genotype data for families from both populations were generated at the same set of marker loci conditional on the two siblings being affected. IBD sharing was estimated using GENEHUNTER [Kruglyak et al., 1996].

Different levels of information content were attained by simulating genotypes at markers with 2, 3, 4, or 10 equally frequent alleles. When our estimation method was applied to less-than-fully-informative IBD sharing data with different levels of information content (Table VI), less informative markers led to greater downward bias in estimates of the $C$ parameters (i.e., lower estimates of expected sharing in ASPs at the disease genes). Furthermore, lower information content led to greater bias in location estimates, with the two disease loci estimated to be further apart than they really were, and thus to lower coverage of the corresponding confidence intervals.

## DISCUSSION

We have introduced and studied the properties of a method for the simultaneous estimation of the locations of two linked disease susceptibility genes. With a sufficiently informative sample, the proposed method provides unbiased estimates with confidence interval coverage near the nominal levels. The method produces better results when the two disease genes are believed to be further apart, when there is higher allele sharing at the two disease genes, and when more ASPs are available for analysis. When the two disease genes are believed to be close together, the method may not be as effective. However, in such cases, other methods may be more appropriate.

<table>
<thead>
<tr>
<th>Model</th>
<th>Initial $\tau$ value</th>
<th>Average estimate</th>
<th>$SD_{emp}$</th>
<th>$SE_{rob}$</th>
<th>Average estimate</th>
<th>$SD_{emp}$</th>
<th>$SE_{rob}$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>50.0</td>
<td>0.211</td>
<td>0.0377</td>
<td>0.0385</td>
<td>49.7</td>
<td>7.09</td>
<td>2.52</td>
<td>6.5</td>
<td>3.8</td>
</tr>
<tr>
<td>A</td>
<td>35.0</td>
<td>0.208</td>
<td>0.0384</td>
<td>0.0384</td>
<td>36.7</td>
<td>5.03</td>
<td>2.58</td>
<td>85.5</td>
<td>0.0</td>
</tr>
<tr>
<td>B</td>
<td>50.0</td>
<td>0.334</td>
<td>0.0396</td>
<td>0.0384</td>
<td>49.9</td>
<td>5.37</td>
<td>1.63</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td>50.0</td>
<td>0.635</td>
<td>0.0302</td>
<td>0.0309</td>
<td>50.0</td>
<td>4.70</td>
<td>0.74</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>C</td>
<td>35.0</td>
<td>0.627</td>
<td>0.0317</td>
<td>0.0308</td>
<td>36.0</td>
<td>0.68</td>
<td>0.76</td>
<td>77.7</td>
<td>0.0</td>
</tr>
<tr>
<td>D</td>
<td>50.0</td>
<td>0.382</td>
<td>0.0386</td>
<td>0.0384</td>
<td>41.6</td>
<td>3.09</td>
<td>1.43</td>
<td>19.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Results are based on 1,000 samples of 500 independent ASPs and a map of 11 markers spaced at 10-cM intervals. $\tau_1 = 35$ and $\tau_2 = 65$ (the resulting $C$'s can be found in the last column of Table I). $SD_{emp}$ = empirical standard deviation of estimates for 1,000 replications in a simulation. $SE_{rob}$ = average robust standard error of estimates for 1,000 replications in a simulation. CI coverage refers to the proportion of CIs for $\tau$ that cover $\tau_1$ or $\tau_2$.

<table>
<thead>
<tr>
<th>True values</th>
<th>No. ASPs</th>
<th>Average estimates</th>
<th>Mean bias</th>
<th>Empirical SD</th>
<th>Average robust SE</th>
<th>“95%” CI coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_1$, $\tau_2$</td>
<td>$C_1$, $C_2$</td>
<td>$\tau_1$</td>
<td>$\tau_2$</td>
<td>$\tau_1$</td>
<td>$\tau_2$</td>
<td>$\tau_1$</td>
</tr>
<tr>
<td>40,60</td>
<td>0.283,0.283</td>
<td>200</td>
<td>39.6</td>
<td>60.4</td>
<td>-0.4</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>39.7</td>
<td>60.4</td>
<td>-0.3</td>
<td>0.4</td>
<td>2.99</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>39.8</td>
<td>60.2</td>
<td>-0.2</td>
<td>0.2</td>
<td>1.88</td>
</tr>
<tr>
<td>35,65</td>
<td>0.254,0.254</td>
<td>200</td>
<td>35.0</td>
<td>65.4</td>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>35.1</td>
<td>65.0</td>
<td>0.1</td>
<td>0.0</td>
<td>2.97</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>35.0</td>
<td>65.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.95</td>
</tr>
<tr>
<td>25,75</td>
<td>0.222,0.222</td>
<td>200</td>
<td>25.2</td>
<td>74.8</td>
<td>0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>25.0</td>
<td>75.1</td>
<td>0.0</td>
<td>0.1</td>
<td>3.33</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>25.0</td>
<td>75.1</td>
<td>0.0</td>
<td>0.1</td>
<td>1.95</td>
</tr>
</tbody>
</table>

Data were generated under model B from Table I at 11 markers spaced at 10-cM interval.

{|TABLE III. Fitting the one-locus model to data generated under a model with two linked disease genes*|
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

*Results are based on 1,000 samples of 500 independent ASPs and a map of 11 markers spaced at 10-cM intervals. $\tau_1 = 35$ and $\tau_2 = 65$ (the resulting $C$'s can be found in the last column of Table I). $SD_{emp}$ = empirical standard deviation of estimates for 1,000 replications in a simulation. $SE_{rob}$ = average robust standard error of estimates for 1,000 replications in a simulation. CI coverage refers to the proportion of CIs for $\tau$ that cover $\tau_1$ or $\tau_2$. |
TABLE V. Estimation of $\tau_1$ and $\tau_2$ with different marker maps (marker densities, number of markers)\textsuperscript{a}

<table>
<thead>
<tr>
<th>$\tau_1$, $\tau_2$ (cM)</th>
<th>$C_1$, $C_2$</th>
<th>Distance between $\tau_1$ and $\tau_2$</th>
<th>No markers (total span of markers in cM)</th>
<th>Marker spacing (cM)</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>40,60</td>
<td>0.283, 0.283</td>
<td>20</td>
<td>11 (0-100)</td>
<td>10</td>
<td>39.7</td>
<td>60.3</td>
<td>-0.3</td>
<td>0.3</td>
<td>2.61</td>
<td>2.61</td>
<td>2.44</td>
<td>2.45</td>
</tr>
<tr>
<td>21 (0-100)</td>
<td>5</td>
<td></td>
<td></td>
<td>2.5</td>
<td>39.9</td>
<td>60.3</td>
<td>-0.1</td>
<td>0.3</td>
<td>2.62</td>
<td>2.47</td>
<td>2.32</td>
<td>2.35</td>
</tr>
<tr>
<td>41 (0-100)</td>
<td>40</td>
<td></td>
<td></td>
<td>10</td>
<td>39.9</td>
<td>60.3</td>
<td>-0.1</td>
<td>0.3</td>
<td>2.84</td>
<td>2.65</td>
<td>2.20</td>
<td>2.22</td>
</tr>
<tr>
<td>30,70</td>
<td>0.235, 0.235</td>
<td>40</td>
<td>11 (0-100)</td>
<td>10</td>
<td>29.8</td>
<td>70.4</td>
<td>-0.2</td>
<td>0.4</td>
<td>2.51</td>
<td>2.58</td>
<td>2.72</td>
<td>2.65</td>
</tr>
<tr>
<td>21 (0-100)</td>
<td>5</td>
<td></td>
<td></td>
<td>2.5</td>
<td>29.8</td>
<td>70.0</td>
<td>-0.2</td>
<td>0.0</td>
<td>2.52</td>
<td>2.46</td>
<td>2.37</td>
<td>2.37</td>
</tr>
<tr>
<td>41 (0-100)</td>
<td>2.5</td>
<td></td>
<td></td>
<td>10</td>
<td>29.8</td>
<td>70.0</td>
<td>-0.2</td>
<td>0.0</td>
<td>2.61</td>
<td>2.57</td>
<td>2.30</td>
<td>2.27</td>
</tr>
<tr>
<td>15,35</td>
<td>0.283, 0.283</td>
<td>40</td>
<td>11 (0-50)</td>
<td>5</td>
<td>14.8</td>
<td>35.0</td>
<td>-0.2</td>
<td>0.0</td>
<td>1.81</td>
<td>1.98</td>
<td>1.90</td>
<td>1.90</td>
</tr>
<tr>
<td>21 (0-22)</td>
<td>10</td>
<td></td>
<td></td>
<td>2</td>
<td>5.9</td>
<td>16.0</td>
<td>-0.1</td>
<td>0.0</td>
<td>1.40</td>
<td>1.42</td>
<td>1.24</td>
<td>1.25</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Based on 1,000 replicates of 500 ASPs. Data generated under model B from Table I.

TABLE VI. Evaluation of our estimation method applied to data from non-fully-informative markers\textsuperscript{a}

<table>
<thead>
<tr>
<th>No. of alleles</th>
<th>Approximate info content</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>92-95%</td>
<td>0.210</td>
<td>0.209</td>
<td>-0.001</td>
<td>-0.002</td>
<td>34.9</td>
<td>65.0</td>
<td>-0.1</td>
<td>0.0</td>
<td>3.65</td>
<td>4.09</td>
<td>3.22</td>
<td>3.40</td>
</tr>
<tr>
<td>4</td>
<td>80-86%</td>
<td>0.200</td>
<td>0.202</td>
<td>-0.011</td>
<td>-0.009</td>
<td>34.8</td>
<td>65.0</td>
<td>-0.2</td>
<td>0.0</td>
<td>3.93</td>
<td>3.69</td>
<td>3.11</td>
<td>3.09</td>
</tr>
<tr>
<td>3</td>
<td>72-79%</td>
<td>0.192</td>
<td>0.194</td>
<td>-0.019</td>
<td>-0.017</td>
<td>34.7</td>
<td>65.1</td>
<td>-0.3</td>
<td>0.1</td>
<td>3.59</td>
<td>3.78</td>
<td>3.06</td>
<td>3.04</td>
</tr>
<tr>
<td>2</td>
<td>52-62%</td>
<td>0.170</td>
<td>0.171</td>
<td>-0.041</td>
<td>-0.040</td>
<td>34.6</td>
<td>65.5</td>
<td>-0.4</td>
<td>0.5</td>
<td>3.95</td>
<td>3.75</td>
<td>2.86</td>
<td>2.84</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Based on 1,000 replicates of 500 ASPs (250 from population 1+250 from population 2, as described in the text). Different levels of information content (2nd column) are achieved by varying the number of equally frequent alleles at each marker (1st column). For all simulations in this table $C_1=C_2=0.211$, $\tau_1=35$, $\tau_2=65$. A map of 11 markers spaced at 10-cM intervals (i.e., markers at 0, 10, ..., 90, 100 cM) was used. Approximate information content across the chromosome was obtained from Genehunter 2.0 and is based on the definition of information content described by Kruglyak et al. [1996].
to be located close to one another, a denser map of markers is recommended, to ensure that there are several markers (3–4) between the two disease genes.

When markers are not fully informative, estimates of the C parameters (expected IBD sharing among ASPs at the disease genes) are biased toward the null. This is because the estimated marker IBD sharing values ($S_{im}$) tend to be lower than the true unknown IBD sharing ($S_{im}$) when there are disease gene(s) linked to the markers, producing a downward bias in estimates of the C’s. The lower C’s are associated with greater bias in the disease gene location estimates, with lower coverage of the corresponding confidence intervals. How much lower the $S_{im}$ values are compared to the $S_{im}$ values, depends on the underlying genetic model (the true $S_{im}$ values), and the level of information content available from the marker data.

In this study, confidence intervals were constructed using asymptotic GEE theory, under which the estimates are normally distributed and the robust variance estimator provides consistent variance estimates. Although for some underlying genetic disease models with sufficient sample size, coverage of asymptotic theory confidence intervals reached the nominal levels, for some situations studied coverage was below the nominal level. In light of these findings, we caution that in the diabetes data application presented in this report, the confidence interval for the second putative disease gene may not have the nominal 95% coverage. The actual coverage may be lower because, in this study, neither the sample size (356 ASPs) nor the estimated effect size of the second gene were large.

Coverage of confidence intervals can be different from the nominal level for a number of reasons: biased parameter estimates (i.e., biased estimates of the $\tau$‘s), biased robust variance estimates, or non-normality of the parameter estimates. Our simulation study indicated that the robust variance estimates of the location estimates tend to be downwardly biased. There was also evidence of lack of normality of the estimates (results not shown). It is not unexpected for the robust estimator to yield downwardly biased variance estimates [e.g., see Pan, 2001]. Methods have been proposed to correct for the bias and the variability [e.g., Mancl and DeRouen, 2001; Pan, 2001; Pan and Wall, 2002] of the robust variance estimator. Such techniques could be applied here, in particular those that adjust for biases of the robust variance estimator. However, the improvements may be modest especially if assumptions of these methods are violated. Confidence interval coverage may be improved by replacing the robust variance estimates by bootstrap estimates of the variances. Confidence interval properties could potentially be further improved in situations when the estimates do not follow the asymptotic normal distribution, by estimating confidence interval endpoints by bootstrap quantiles.

Using a large number of markers in the analysis can lower the confidence interval coverage. Asymptotic properties of the GEE method are approached by increasing the number of independent sampling units in the analysis, which in this case is accomplished by increasing the number of independent affected sib pairs. By increasing the number of markers without increasing the number of ASPs, the clusters comprised of dependent data increase in size without proportionally increasing the effective sample size (i.e., number of independent units). This leads to a greater downward bias in the variance estimates provided by the robust variance estimator, and thus to less than nominal confidence interval coverage. This may also be exacerbated by the fact that we did not optimally model Cov($S_{ij}$|\Phi). We assumed an independence correlation structure, which may decrease efficiency of parameter estimates, and empirically estimated the variance in IBD sharing at each marker independent of the other markers, which essentially amounts to allowing a new parameter for the variance at each marker. As the number of markers increases, the number of unknown parameters estimated from the data increases, requiring a larger sample size to achieve asymptotic properties.

Although confidence interval coverage can decrease as the number of markers increases, there are benefits resulting from the use of additional markers. Bias of parameter estimates may be reduced and the true variances of estimates, as estimated by the sample variances from all the replicates in a simulation, generally decrease. Thus, more accurate and more precise estimates can be achieved when more markers are used in the analysis. Greater benefits from using additional markers would be seen if the marker data were not fully informative. Because we observed that using a large number of markers may lead to greater bias in the robust variance estimates and lower confidence interval coverage,
we conclude that when a large number of markers are available for analysis, greater care must be taken in constructing the confidence intervals. Also, as the number of linked markers increases, modeling the variances and correlations in IBD sharing at these markers may become more beneficial.

We observed that fitting the one-locus model introduced by Liang et al. [2001] can produce misleading results when there are two linked disease genes in the region. The single disease gene location is likely to be estimated at an incorrect position between the two true disease gene locations while the corresponding effect size tends to be over-estimated. Furthermore, the standard error of the location estimate tends to be biased down, leading to a confidence interval for a disease gene location that frequently does not cover either of the two true locations. This is not surprising, since the GEE approach is known to provide consistent estimates of the parameters under mild regularity conditions, provided that the mean function has been correctly specified, which is not the case if the one-locus mean function is postulated for data generated under a model with two linked disease loci. It should also be noted that the situation presented here, with the two loci having equal effect size, is a worst-case scenario in terms of how different the estimates based on the two models will be. If one of the genes had a substantially higher effect than the other, as is the case with the diabetes data application we present, the stronger gene may be localized with reasonable precision by fitting the one-locus model, although the estimate would be biased in the direction of the second gene.

Since fitting an incorrect model can lead to very biased estimates with highly anti-conservative confidence intervals, it is important to decide whether a one-locus or two-locus model is more appropriate. We are developing and evaluating test statistics that would assist in making this decision [Biernacka and Bull, 2003]. Farrall [1997] described a likelihood approach for testing the independent support of a putative susceptibility gene that maps close to a previously established gene. In his approach, it was assumed that an "anchor" gene has already been mapped, and evidence for linkage at linked loci was evaluated taking into account the first mapped gene. However, as our results and the results of other studies [Hauser et al., 2003] have shown, localization of the first gene may be incorrect if the presence of other linked disease genes is not taken into account. "Conditional approaches" for mapping a second disease gene linked to an already mapped disease gene may be suitable in situations when the first gene has a much stronger effect than the second linked gene, as is the case for the diabetes data. However, if both genes have a similar effect size (or the difference in effect size is not large), these approaches would probably be inferior to an approach that maps the two genes simultaneously. In such cases, our methods can help localize the two genes more precisely.

Under the same assumptions that were used to derive the two-locus model, the method can be easily extended to allow estimation of three or more linked disease genes. However, there are likely to be more computational problems as the number of linked disease genes is increased.

Although localization of two linked disease genes (i.e., estimation of $\tau_1$ and $\tau_2$) is the primary purpose of the method introduced in this report, $C_1$ and $C_2$, i.e., the excess IBD sharing in ASPs at the two disease genes, are also estimated. In the one-locus model, excess IBD sharing at a single disease gene on a chromosome can be interpreted as the effect size of that gene. This effect size is 0 under the null hypothesis of no disease genes linked to the region. However, this simple "effect size" interpretation does not apply to $C_1$ and $C_2$ in the two-locus model. When two disease genes are linked, expected IBD sharing between affected relatives at any one of these genes increases as a result of effects of both genes. Therefore, for example, $C_1$ does not represent the effect of gene 1 alone, because even in the absence of any effect of gene 1, $C_1$ will increase as the distance between gene 1 and gene 2 decreases, due to the effect of gene 2. An alternative parameterization of the two-locus model is described in the Appendix, which simplifies the mean function formula, and is more conducive to a "gene effect size" interpretation of the parameters. With this alternative formulation of the mean function, a GEE approach equivalent to the one described in this study could be used to estimate the alternative set of parameters.

In conclusion, we have introduced, applied, and evaluated a method for estimating the locations of two linked disease susceptibility genes. This method can provide unbiased estimates of disease gene locations, while avoiding multiple testing...
problems by looking at all markers jointly. Because this method provides confidence intervals for the two gene locations, it can help define region(s) for further studies aimed at fine-mapping of the trait loci.

ACKNOWLEDGMENTS

We gratefully acknowledge the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory for allowing us to use their diabetes genome scan data (data version 2.0) [Mein et al., 1998]. Those who carried out the original analysis and collection of this data bear no responsibility for our further analysis and interpretation of it. This research was supported by an Ontario Student Opportunity Trust Fund Award administered by the Samuel Lunenfeld Research Institute (J.M.B) and by funding from the Mathematics for Information Technology and Complex Systems (MITACS) Network of Centres of Excellence (NCE-Canada). J.M.B. was also supported by a Natural Sciences and Engineering Research Council of Canada graduate award and an Ontario Graduate Scholarships and Engineering Research Council of Canada Network of Centres of Excellence (NCE-Canada). S.B.B. is a Senior Investigator of the Canadian Institutes of Health Research.

REFERENCES


Hauser ER, Bass M, Martin ER. 2003. Identification of gene locations from maximum likelihood ASP linkage analysis: are there features of the lod score curve that distinguish regions with two loci? Am J Hum Genet 73(Suppl):615.


APPENDIX

DERIVATION OF E[S(t)|Φ]

Let τ1 and τ2 be the locations of two linked trait loci (τ1 < τ2). Let S(t) be the number of alleles (0, 1, or 2) shared IBD by an ASP at locus τ, and Φ denote the event that both sibs are affected. First note that

\[ E[S(t)|Φ] = Pr(S(t) = 1|Φ) + 2 Pr(S(t) = 2|Φ) \]
Also,

\[
\Pr(S(t) = j|\Phi) = \frac{\Pr(S(t) = j, \Phi)}{\Pr(\Phi)}
\]

\[
= \sum_{k=0}^{2} \sum_{l=0}^{2} \Pr(S(t) = j, S(\tau_1) = k, S(\tau_2) = l) \Pr(\Phi)
\]

\[
= \sum_{k=0}^{2} \sum_{l=0}^{2} \Pr(\Phi|S(t) = j, S(\tau_1) = k, S(\tau_2) = l) \Pr(S(t) = j, S(\tau_1) = k, S(\tau_2) = l) \Pr(\Phi)
\]

\[
= \sum_{k=0}^{2} \sum_{l=0}^{2} \Pr(\Phi|S(\tau_1) = k, S(\tau_2) = l) \Pr(S(\tau_2) = j|S(\tau_1) = k, S(\tau_2) = l) \Pr(S(\tau_1) = k, S(\tau_2) = l)
\]

\[
= \sum_{k=0}^{2} \sum_{l=0}^{2} \Pr(S(\tau_1) = k, S(\tau_2) = l) \Pr(S(t) = j|S(\tau_1) = k, S(\tau_2) = l)
\]

\[
= \sum_{k=0}^{2} \sum_{l=0}^{2} \Pr(S(t) = j|S(\tau_1) = k) \Pr(S(\tau_1) = k|\Phi)
\]

As the above formula shows, in order to calculate the probabilities \(\Pr(S(t) = 1|\Phi)\) and \(\Pr(S(t) = 2|\Phi)\), we need conditional probabilities of a sib pair sharing \(i\) alleles IBD at \(t\) given that they share \(j\) alleles IBD at \(\tau_1\). The matrix of conditional sharing probabilities in sibpairs \(\Pr(S_A = i|S_B = j)\) for two linked loci A and B was given by Haseman and Elston [1972]. Using those probabilities, after simplification, it can be shown that

\[
E[S(t)|\Phi] = 1 + (2\Psi_1 - 1)C_1
\]

where \(C_1 = E[S(\tau_1)|\Phi] - 1\) and \(\Psi_1 = \Psi_{i,\tau_1} = 0_{i,\tau_1}^2 + \sigma_0^2 (1 - \theta_{i,\tau_1})^2\).

**Case II: \(\tau_1 < \tau_2 < t\)**. Follows similarly to case I.

**Case III: \(t < \tau_1 < \tau_2\)**

Under the assumption of no interference:

\[
\Pr(S(t) = j|S(\tau_1) = k, S(\tau_2) = l)
\]

\[
= \frac{\Pr(S(t) = j, S(\tau_1) = k, S(\tau_2) = l)}{\Pr(S(\tau_2) = l|S(\tau_1) = k)}
\]

Therefore,

\[
E(S(t)|\Phi) = \{\Pr(S(t) = 1|\Phi) + 2\Pr(S(t) = 2|\Phi)\}
\]

\[
= \sum_{k=0}^{2} \sum_{l=0}^{2} \{\Pr(S(\tau_1) = k, S(\tau_2) = l) \times \Pr(S(t) = 1|S(\tau_1) = k, S(\tau_2) = l)\}
\]

\[
+ 2 \sum_{k=0}^{2} \sum_{l=0}^{2} \{\Pr(S(\tau_1) = k, S(\tau_2) = l) \times \Pr(S(t) = 2|S(\tau_1) = k, S(\tau_2) = l)\}
\]

\[
= \sum_{k=0}^{2} \sum_{l=0}^{2} g_{kl} \times \Pr(S(\tau_2) = l|S(t) = 1) \Pr(S(t) = 1|S(\tau_1) = k)
\]

\[
+ \sum_{k=0}^{2} \sum_{l=0}^{2} g_{kl} \times \Pr(S(\tau_2) = l|S(t) = 2) \Pr(S(t) = 2|S(\tau_1) = k)
\]

where \(g_{kl} = \Pr(S(\tau_1) = k, S(\tau_2) = l|\Phi)\).
By substituting in expressions for the conditional sharing probabilities, and using the properties:

\[
\sum_{k=0}^{2} \sum_{l=0}^{2} g_{kl} = 1,
\]

\[
E(S(\tau_1)|\Phi) = Pr(S(\tau_1) = 1|\Phi) + 2 Pr(S(\tau_1) = 2|\Phi)
= \sum_{l=0}^{2} Pr(S(\tau_1) = 1, S(\tau_2) = l|\Phi)
+ 2 \sum_{l=0}^{2} Pr(S(\tau_1) = 2, S(\tau_2) = l|\Phi)
= \sum_{l=0}^{2} g_{kl} + 2 \sum_{l=0}^{2} g_{kl}
\]

and

\[
E(S(\tau_2)|\Phi) = Pr(S(\tau_2) = 1|\Phi) + 2 Pr(S(\tau_2) = 2|\Phi)
= \sum_{k=0}^{2} Pr(S(\tau_1) = k, S(\tau_2) = 1|\Phi)
+ 2 \sum_{k=0}^{2} Pr(S(\tau_1) = k, S(\tau_2) = 2|\Phi)
= \sum_{k=0}^{2} g_{k1} + 2 \sum_{k=0}^{2} g_{k2}
\]

it can be shown that:

\[
E(S(t)|\Phi) = \frac{\Psi_3(\Psi_2 - 1)(2\Psi_1 - 1)}{(1 - \Psi_1 - \Psi_2 + 2\Psi_1\Psi_2)(2\Psi_1\Psi_2 - \Psi_1 - \Psi_2)} \times (C_1 + 1)
+ \frac{\Psi_1(\Psi_1 - 1)(2\Psi_2 - 1)}{(1 - \Psi_1 - \Psi_2 + 2\Psi_1\Psi_2)(2\Psi_1\Psi_2 - \Psi_1 - \Psi_2)} \times (C_2 + 1)
- \frac{(2\Psi_1 - 1)(2\Psi_2 - 1)}{1 + (2\Psi_1 - 1)(2\Psi_2 - 1)}
= 1 + \frac{(2\Psi_1 - 1)(1 - (2\Psi_1 - 1)^2)}{1 - (2\Psi_3 - 1)^2} C_1
+ \frac{(2\Psi_2 - 1)(1 - (2\Psi_1 - 1)^2)}{1 - (2\Psi_3 - 1)^2} C_2
= 1 + \frac{(2\Psi_1 - 1)(1 - (2\Psi_1 - 1)^2)}{1 - (2\Psi_3 - 1)^2} C_1
+ \frac{(2\Psi_2 - 1)(1 - (2\Psi_1 - 1)^2)}{1 - (2\Psi_3 - 1)^2} C_2
\]

where \(\Psi_1 = \Psi_{1,\tau_1}, \Psi_2 = \Psi_{1,\tau_2}\) and \(\Psi_3 = \Psi_{1,\tau_2}^2\).

\(\text{Var}(S(t)|\Phi)\)

The variances in IBD sharing at the two disease loci are

\[
\text{Var}(S(\tau_1)|\Phi) = C_1(1 - C_1) + 2 Pr(S(\tau_1) = 0|\Phi) \\text{and} \\text{Var}(S(\tau_2)|\Phi) = C_2(1 - C_2) + 2 Pr(S(\tau_2) = 0|\Phi).
\]

The variance at a fully informative marker \(t\) depends on the location of \(t\) in relation to the two disease loci.

For \(t < \tau_1 < \tau_2\) : \(\text{Var}(S(t)|\Phi)\)

\[
= (2\Psi_1 - 1)[(\text{Var}(S(\tau_1)|\Phi) - 0.5] + 0.5
\]

For \(\tau_1 < t < \tau_2\) : \(\text{Var}(S(t)|\Phi)\)

\[
= (2\Psi_2 - 1)[(\text{Var}(S(\tau_2)|\Phi) - 0.5] + 0.5
\]

For \(\tau_1 < t < \tau_2\) : \(\text{Var}(S(t)|\Phi)\)

\[
= E(S(t)|\Phi) - [E(S(t)|\Phi)]^2 + 2 Pr(S(t) = 2|\Phi).
\]

\(E(S(t)|\Phi)\) is given above, and is a function \(C_1, C_2, \tau_1\) and \(\tau_2\). \(Pr(S(t) = 2|\Phi)\) is a function of the distance between \(t\) and \(\tau_1\) and \(\tau_2\), as well as the nine parameters \(g_{ij} = Pr(S(\tau_1) = i, S(\tau_2) = j|\Phi)\) \(i,j = 0,1,2(\sum_{i=0}^{2}\sum_{j=0}^{2}g_{ij} = 1)\), which depend on the disease model through penetrances and disease allele frequencies at the disease loci, and on the distance between \(\tau_1\) and \(\tau_2\), as follows,

\[
Pr(S(t) = 2|\Phi) = \frac{(1 - \Psi_1)^2(1 - \Psi_2)^2g_{00}}{\Psi_3^2} + \frac{(1 - \Psi_1)^2\Psi_2(1 - \Psi_2)g_{01}}{\Psi_3(1 - \Psi_3)} + \frac{(1 - \Psi_1)^2\Psi_2^2g_{02}}{(1 - \Psi_3)^2} + \frac{\Psi_1(1 - \Psi_1)(1 - \Psi_2)^2g_{10}}{(1 - \Psi_3)^2} + \frac{\Psi_1(1 - \Psi_1)\Psi_2^2g_{11}}{(1 - \Psi_3)^2} + \frac{\Psi_1^2\Psi_2(1 - \Psi_2)g_{12}}{\Psi_3^2} + \frac{\Psi_1^2\Psi_2^2g_{20}}{(1 - \Psi_3)^2} + \frac{\Psi_1^2\Psi_2^2(1 - \Psi_2)g_{21}}{\Psi_3^2} + \frac{\Psi_1^2\Psi_2^2g_{22}}{\Psi_3^2}
\]

\(\text{Cov}(S(t)|\Phi)\)

In order to solve the GEE (equation 1), one needs to formulate \(\text{Cov}(S(t)|\Phi)\) as a function of the mean, \(\mu_2\), and possibly other nuisance parameters. If \(\text{Cov}(S(t)|\Phi)\) is unknown, we may assume a working covariance matrix for \(S(t)\). A solution to the GEEs provides consistent estimates of the
parameters $\delta_2$ even when the correlation structure for the data is incorrectly specified, and the robust sandwich estimator [White, 1982; Liang and Zeger, 1986] gives consistent estimates of the variances of parameter estimates.

In this implementation of the two-locus model, we assume an independence correlation structure by setting $R_s(\alpha)$ equal to the identity matrix. That is, we assume that the correlation of IBD sharing at any two markers for the same sib pair is 0, and correlation of IBD sharing across sibling pairs from the same family is also 0, and rely on the empirical component of the robust sandwich estimator to capture the correlation information. Therefore, Cov$(S_j^r|\Phi)$ is a $M \times M$ diagonal matrix with Var$(S_j^r(t_m)|\Phi)$, $m = 1, \ldots, M$ on the diagonal. We estimate Var$(S_j^r(t_m)|\Phi)$ empirically by:

$$\text{V} \text{a} \text{r}(S_j^r(t_m)|\Phi) = \frac{1}{n-1} \sum_{i=1}^{n} (S_j^r(t_m) - S^*(t_m))^2,$$

where $S^*(t_m) = \frac{\sum_{i=1}^{n} S_j^r(t_m)}{n}$, $n=$number of ASPs.

Under a one-locus model, our simulations have shown that empirically estimating variances of $S^*$ at the markers led to results comparable to those obtained by modelling Var$(S_j(t_m)|\Phi)$ using an approximate variance function implemented in Genefinder [Liang et al., 2001].

AN ALTERNATIVE PARAMETERIZATION

Consider the following re-parameterization in the expression for $\mu_2(t; \delta_2)$:

$$C_1 = C_1^* + C_2^e^{-0.04(t_2 - \tau_1)} \quad \text{and} \quad C_2 = C_2^* + C_1^e^{-0.04(t_2 - \tau_1)}.$$

That is:

$$C_1^* = \frac{C_1 - C_2e^{-0.04(t_2 - \tau_1)}}{1 - e^{-0.08(t_2 - \tau_1)}} \quad \text{and} \quad C_2^* = \frac{C_2 - C_1e^{-0.04(t_2 - \tau_1)}}{1 - e^{-0.08(t_2 - \tau_1)}}.$$

Then it can be shown that,

$$\mu_2(t) = E(S(t)|\Phi) = 1 + C_1^*e^{-0.04[t - \tau_1]} + C_2^*e^{-0.04[t - \tau_2]} \quad \text{for all } t.$$

Note that if $C_1^* = 0$ then $\mu_2(t) = \mu_1(t) = E(S(t)|\Phi) = 1 + C_2^*e^{-0.04[t - \tau_1]}$. Similarly, if $C_2^* = 0$ then $\mu_2(t) = \mu_1(t) = E(S(t)|\Phi) = 1 + C_1^*e^{-0.04[t - \tau_1]}$.

That is, the two locus model parameterized by $\delta^* = (C_1^*, C_2^*, \tau_1, \tau_2)$ reduces to the one locus model if either of the $C^*$ parameters equals 0. This implies that the $C^*$ parameters represent the excess sharing at a location due to the gene at that location, and can be seen as the “effect size” of that particular gene, whereas the original $C$ parameters ($C_1$ and $C_2$) represent the amount of excess IBD sharing at each of the two disease gene loci, respectively, which is increased by effects due to both disease genes.

Although the interpretation of $C_1$ and $C_2$ is clear (i.e., one less than the expected IBD sharing at the locations of the two disease genes, $\tau_1$ and $\tau_2$), interpretation of $C_1^*$ and $C_2^*$ is not as obvious. Empirical results (not shown) suggest that when the two-locus disease model is additive on the penetrance scale, $C_1^*$ and $C_2^*$ equal the excess IBD sharing at two disease genes under a model with the same penetrance matrix and allele frequencies, but with the two disease genes unlinked. In this case, $C_1^*$ and $C_2^*$ can be interpreted as the effect on IBD sharing due to gene 1 and gene 2, respectively. For other non-additive models, interpretation of $C_1^*$ and $C_2^*$ is not so straightforward.