Robust Variance-Components Approach for Assessing Genetic Linkage in Pedigrees

Christopher I. Amos

Department of Epidemiology, University of Texas, M. D. Anderson Cancer Center, and Genetic Studies Section, N.I.A.M.S., National Institutes of Health, Bethesda

Summary

To assess evidence for genetic linkage from pedigrees, I developed a limited variance-components approach. In this method, variability among trait observations from individuals within pedigrees is expressed in terms of fixed effects from covariates and effects due to an unobservable trait-affecting major locus, random polygenic effects, and residual nongenetic variance. The effect attributable to a locus linked to a marker is a function of the additive and dominance components of variance of the locus, the recombination fraction, and the proportion of genes identical by descent at the marker locus for each pair of sibs. For unlinked loci, the polygenic variance component depends only on the relationship between the relative pair. Parameters can be estimated by either maximum-likelihood methods or quasi-likelihood methods. The forms of quasi-likelihood estimators are provided. Hypothesis tests derived from the maximum-likelihood approach are constructed by appeal to asymptotic theory. A simulation study showed that the size of likelihood-ratio tests was appropriate but that the monogenic component of variance was generally underestimated by the likelihood approach.

Introduction

The first methods (Penrose 1938; Brues 1950) used to investigate genetic linkage between a quantitative trait and a marker assessed correlation between pairs of relatives and did not require explicit assumptions about the genetic model describing interindividual trait variability. Once modern computers became available these approaches were largely abandoned because they have less power compared with model-dependent maximum-likelihood methods for diseases with clear inheritance patterns. However, in light of the complexity of the processes that determine the interindividual variability of most quantitative traits, methods that are not dependent on specifying a correct genetic model have become popular lately (MacCluer 1989; Jeunemaitre et al. 1992). For quantitative traits, models based on regression analysis (Haseman and Elston 1972; Carey and Williamson 1991; Fulker et al. 1991) and variance-components analysis (Chevalet et al. 1984; Goldgar 1990) have been developed. For randomly selected pedigrees, the Haseman-Elston approach has been shown to perform well when compared with maximum-likelihood methods based on knowledge of the correct genetic model (Kammerer and MacCluer 1985; Demenais and Amos 1989). Recently the Haseman-Elston method was extended to allow analysis of all pedigree members (Amos and Elston 1989; Olson and Wijsman 1993). An intricate adaptation of the quasi-likelihood method extended the Haseman-Elston methods so that information from various pedigree structures could be optimally combined (Olson and Wijsman 1993).

Variance-components approaches to detecting genetic linkage have been mainly developed for the study of experimental organisms. Jayakar (1970) developed a variance-components approach that is applicable for the offspring from matings in which the phase is known. Similarly, Lander and Botstein (1989) developed an approach that permits inclusion of multiple markers, with the assumption that the trait values are normally distributed in the parental and progeny generations. Goldgar (1990) developed a variance-components approach for analyzing nonexperimental sibships that has greater power than the Haseman-Elston app-
proach when data from several linked marker loci are available, but this approach assumes that several genetic factors from a chromosomal region influence the trait. The properties of Goldgar's approach and its power have been extensively studied (Schork 1993). Here, I develop a mixed-effects variance-components approach for evaluating covariate effects, as well as evidence for genetic linkage to a single trait-affecting locus for data from pedigrees. This method can more easily accommodate pedigree data than does the Haseman-Elston approach, can provide estimates of heritability, and can provide an estimate of the recombination fraction between a marker and a trait-affecting locus. Estimators based on maximum-likelihood models or quasi-likelihood methods are developed. The latter approach has the advantage of not requiring multivariate normality assumptions. The performance of the variance-components approach, fitted by maximum-likelihood methods, is assessed in a simulation study based on analysis of sibships.

The Genetic Model

If \( X_i \) represents the trait value of the \( i \)th relative and \( z_{ik} \) is the \( k \)th covariate measurement (of \( s \) covariates) on this individual, then a general model describing the trait values is

\[
X_i = \mu + g_i + G_i + \sum_{k=1}^{s} \beta_k z_{ik} + e_i,
\]

where \( \mu \) is the overall mean, and a fixed, unobserved major gene component \( g_i \), with alleles A and B, affects the trait by effects of

\[
g_i = \begin{cases} 
  a & \text{if individual } i \text{ has unobserved genotype } AA \\
  d & \text{if individual } i \text{ has unobserved genotype } AB \\
  -a & \text{if individual } i \text{ has unobserved genotype } BB.
\end{cases}
\]

Let \( \beta_k \) be a covariate effect uncorrelated with the genetic factors, and let \( e_i \) be residual variability uncorrelated with the genetic factors and covariates. \( G_i \) is a random polygenic effect. Without loss of generality, \( E(g_i) = E(e_i) = E(G_i) = 0 \) can be assumed, since any average effects can be incorporated into the overall mean of the data. Under these assumptions, the first two moments describing the data in outbred pedigrees are (Lange et al. 1976)

\[
E(X_i) = \mu + \sum_{k=1}^{s} \beta_k z_{ik} \quad (1a)
\]

and

\[
\text{Cov}(X_i, X_j) = \begin{cases} 
  \sigma^2_g + \sigma^2_d + \sigma^2_G + \sigma^2_e & \text{if } i = j \\
  \Phi_p \sigma^2_g + \Delta_p \sigma^2_d + \Phi_q \sigma^2_G & \text{if } i \neq j
\end{cases}
\]

(1b)

where \( \sigma^2_G \) and \( \sigma^2_d \) are, respectively, the polygenic and residual components of variance, while \( \sigma^2_g = 2pq[p(1-p)]^2 \) and \( \sigma^2_d = 4pq^2d^2 \), where \( p \) is the gene frequency of A, and \( q \) of B. \( \Phi_p \) is the coefficient of relationship between pairs of individuals, while \( \Delta_p \) is the probability that a pair shares both alleles at the major locus identical by descent, based only on their genetic relationship. Values for \( \Phi_p \) and \( \Delta_p \) for common pairs of relatives are well-known, being, e.g., 1/2 and 1/4, respectively, for sibs, and 1/2 and 0, respectively, for other outbred relationships, where \( k \) is the degree of relationship. Under this model, major-gene and polygenic effects are completely confounded without consideration of higher moments.

Now suppose that identity-by-descent (\( \pi_{ij} \)) sharing, as assessed by marker typings for pairs of individuals (and possibly their relatives), is observable for the major locus. Data for such a model result from study of a candidate gene in which the specific genetic variations influencing a trait of interest are unknown, but highly polymorphic markers are available within or very near the candidate gene. In this case, a more precise formulation of the variation among pedigree members is

\[
E(X_i | \pi_{ij}) = \mu + \sum_{k=1}^{s} \beta_k z_{ik} \quad (2a)
\]

and

\[
\text{Cov}(X_i, X_j | \pi_{ij}) = \begin{cases} 
  \sigma^2_g + \sigma^2_d + \sigma^2_G + \sigma^2_e & \text{if } i = j \\
  \pi_{ij} \sigma^2_g + \Delta_{ij} \sigma^2_d + \Phi_q \sigma^2_G & \text{if } i \neq j
\end{cases}
\]

(2b)

where \( \pi_{ij} \) is the proportion of genes that are identical by descent, and \( \Delta_{ij} \) is the probability that the pair shares both genes identical by descent for the major locus. If observable polymorphism in a candidate gene affects interindividual trait variability, the first moment of the model can be modified to


\[ E(X_i | \pi_{ij}) = \mu + \sum_{k=1}^{p} \beta_k z_{i,k} + h_i, \]

where \( h_i \) is a fixed effect attributable to each of the possible observed genotypes of the candidate locus. For example, the actual trait-influencing genotypes \( g \), might include some of those detectable \( (h_i) \) by a particular assay. As in standard analysis-of-covariance approaches, the fixed effects can be modeled by constructing dummy variables to contrast mean trait levels of the various genotypes. For example, one may constrain the most common genotype to have mean effect 0 so that other genotypic effects then represent deviations from this genotype. This model is plausible if the alleles of the candidate locus contribute to some, but not all, of the variability of the locus, so that the detected genotypes directly affect the trait of interest and also identify, in related individuals, additional variability from other polymorphisms within the same gene (or possibly either regulatory sequences in the region or other tightly linked loci).

### Extension to a Linked Marker

Information for assessing genetic linkage usually comes from markers that have no direct effect on the trait of interest. Equation (1) can be extended to include data from a linked marker by considering the cosegregation of a trait and linked marker alleles. A previous approach to extending the Haseman-Elston analysis of outbred pedigree members (Amos and Elston 1989) can be readily adapted for this purpose. The Haseman-Elston approach regresses squared pair differences in the measurements of a trait on identity by descent for a marker locus. When covariate effects are not considered, the major genetic covariance among pairs of individuals is thus

\[
E(X_{ij} - X_j | \pi_{ij})^2 = E(X_i^2) + E(X_j^2) - 2E(X_iX_j | \pi_{ij})
= 2 \text{Var}(X_i) - 2 \text{Cov}(X_i, X_j | \pi_{ij}),
\]

if \( E(X_i^2) = E(X_j^2) \), where \( \pi_{ij} \) represents identity-by-descent sharing for the marker locus, summarized by the proportion of genes identical by descent \( (\pi_{ij}) \) and \( \Delta_{ij} \), the probability that the pair shares two genes identical by descent at the marker locus. Provided the genetic and environmental effects act independently, inclusion of covariate effects does not affect these results. Therefore, covariate effects will not be considered in the following discussion, since they may be incorporated by

Table 1

<table>
<thead>
<tr>
<th>Relative Pair</th>
<th>Component of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sibs</td>
<td>( \theta(1-\theta)^2) ( \pi_{ij} ) ( \sigma_e^2 )</td>
</tr>
<tr>
<td>Half-sibs</td>
<td>( \theta(1-\theta)^2) ( \pi_{ij} ) ( \sigma_e^2 )</td>
</tr>
<tr>
<td>Avuncular</td>
<td>( \theta(1-\theta)^2) ( \pi_{ij} ) ( \sigma_e^2 )</td>
</tr>
<tr>
<td>Grandparental</td>
<td>( \theta(1-\theta)^2) ( \pi_{ij} ) ( \sigma_e^2 )</td>
</tr>
<tr>
<td>First cousin</td>
<td>( \theta(1-\theta)^2) ( \pi_{ij} ) ( \sigma_e^2 )</td>
</tr>
</tbody>
</table>

factoring the likelihood of the data (Corbeil and Searle 1976). Generally, then, the covariance among pairs of individuals in a pedigree can be expressed as

\[
\text{Cov}(X_i, X_j | \pi_{ij}) = \begin{cases} 
\sigma_e^2 + \sigma_g^2 + \sigma_c^2 & \text{if } i = j \\
\theta(\pi_{ij})\sigma_e^2 + g(\theta, \Delta_{ij})\sigma_g^2 + \Phi_{ij}\sigma_c^2 & \text{if } i \neq j
\end{cases}
\]

Table 1 shows the values of \( \theta, \pi_{ij}, \sigma_g^2 \) associated with the additive major-gene component derived from tables 1 and 2 of Amos and Elston (1989) by equation (3). Generally, the dominance component of variance is negligible and cannot be estimated from the covariances among unilineal relatives. However, for sibs, inclusion of a dominance component of variance yields

\[
\text{Cov}(X_i, X_j | \pi_{ij}) = 2\theta(1-\theta)\sigma_e^2 + (1-2\theta)^2 \sigma_g^2 + \left( (1-2\theta)^2 \sigma_e^2 - (1-2\theta)^4 \sigma_g^2 \right) \pi_{ij} + (1-2\theta)^4 \Delta_{ij} \sigma_e^2.
\]

If \( \Delta_{ij} \) is not used in the analysis, then application of equation (3) and results from Amos et al. (1989) yield

\[
\text{Cov}(X_i, X_j | \pi_{ij}) = \left[ \frac{1}{2} + (1-2\theta)^2 \left( \pi_{ij} - \frac{1}{2} \right) \right] \sigma_e^2 - \frac{1}{4} \sigma_g^2,
\]

provided that the parental marker genotypes can be deduced from data.

In addition to the monogenic component of variance, polygenic and residual components of variability need to be incorporated into the analysis. The components of nonmonogenic variance for pairs of individuals can be derived from equation (1b) and the coefficient of relationship, \( \Phi_{ij} \). For instance, for an uninformative marker, \( \text{Cov}(X_i, X_j | \pi_{ij}) = \Phi_{ij}\sigma_e^2 + \Delta_{ij} \sigma_g^2 \). For an un-
linked marker, \( E[\text{Cov}(X_i, X_j|\pi_i)] = \Phi_i \sigma^2 + \Delta_i \sigma^2 \).
Failing to include a polygenic component of variance can, therefore, be expected to lead to false-positive evidence for linkage, which reflects the confounding of \( \sigma^2 \) and \( \sigma^2 \), when an unlinked marker locus is studied. Similarly, data from only one type of relative pair will not yield unique estimates of the four parameters \( \sigma^2, \sigma^2, \sigma^2, \) and \( \theta \): the parameters \( \sigma^2 \) and \( \theta \) are confounded with \( \sigma^2 \) and \( \sigma^2 \) in the covariance expression for pairs of sibs. When data from several types of relative pairs are available, estimating \( \theta \) is possible because \( \theta \) and \( \sigma^2 \) are no longer completely confounded. Data from only sibs and half-sibs will not yield unique estimates of \( \theta \) and \( \sigma^2 \), as these parameters remain confounded for these pairs. When data from only one type of relative pair are available, estimation and hypothesis tests for linkage can be conducted assuming \( \theta = 0 \).

**Estimation**

Parameters can be estimated by maximum-likelihood methods or by estimating-equation approaches. For the maximum-likelihood approaches, assume that \( R \) pedigrees are sampled, each having \( n_i \) members. From the results of Fisher (1918), the trait values of relatives (when the major-gene component of variance is neglected) can be expected to be multivariate normally distributed. Let \( E(X_i) = \mu \), with covariances \( V \), as given by equation (2), table 1, and equations (4) or (5) for sib pairs, if a major genetic dominance component is included. When multivariate normality is assumed, the log likelihood of the data is

\[
\log(L) = c - \frac{1}{2} \sum_{r=1}^{R} \log(\det(V_r)) - \frac{1}{2} \sum_{r=1}^{R} (X_r - \mu_r)'V_r^{-1}(X_r - \mu_r).
\]

Estimating the parameters requires inversion of \( R \) matrices of size \( n_i \) each. The likelihood function can be optimized with available programs (MAXFUN 1992) or, e.g., by the method of scoring (Lange et al. 1976). The hypothesis that the variance components are different from 0 can be tested by asymptotic theory. Standard \( \chi^2 \) tests can be constructed by comparing \(-2\) times the log-likelihood ratio under an unrestricted model (in which \( \sigma^2, \sigma^2, \sigma^2, \) and possibly \( \theta \) are estimated) with a model with constraints imposed by the null hypothesis (\( \sigma^2 \) is constrained to equal 0) with a \( \chi^2 \) distribution. For \( \chi^2 \) tests, the number of df is determined by the number of imposed constraints. It should generally be possible to identify both \( \theta \) and the monogenic variance components when the data contain several types of relative pairs. However, \( \sigma^2 \) cannot be identified when \( \sigma^2 = 0.5 \), and, similarly, \( \theta \) cannot be identified when \( \sigma^2 = 0 \). Thus, a preliminary test for linkage should be constructed by assuming that \( \theta = 0 \) (or some other small value). The estimate of \( \sigma^2 \), so derived, is downwardly biased if \( \theta > 0 \). Variability attributable to a major gene will result in a platykurtotic distribution. I have assessed the effects of assuming multivariate normality in the presence of major-gene segregation, in the section “Simulation Studies.”

**Generalized Estimating-Equation Approaches**

Because one must make an assumption about the distribution of the data to apply maximum-likelihood methods, using more robust estimation procedures seems beneficial. In principle one can apply the generalized estimating-equation approach (Liang and Zeger 1986; McCullagh and Nelder 1989), as recently extended (Prentice and Zhao 1991), to allow simultaneous estimation of structural parameters affecting the means and covariances of multivariate data. Assume that \( R \) family structures are sampled. Let \( \beta = [\beta_1, \beta_2, \ldots, \beta_p] \) (i.e., parameters of the fixed effects) and \( \alpha = [\theta, \sigma^2, \sigma^2, \sigma^2, \sigma^2] \) (i.e., parameters of the random effects). Let \( S_k \) be the matrix of empirical variances and covariances in the \( k \)th family, given the overall mean and any covariate values. \( Y_r \) is an \((n_i+n_i^2) \times 1\) matrix consisting of \( n_i \) trait observations, \( X_i \), followed by \( s_r = \text{vec}(S_k) \). \( T_r \) is an \((n_i+n_i^2) \times 1\) matrix composed of \( \mu \), followed by \( v_r = \text{vec}(V_r) \). To estimate the first and second moments, one can solve the estimating equations

\[
R^{-1/2} \sum_{r=1}^{R} D_r' W_r^{-1} f_r = 0,
\]

where \( f_r = (Y_r - T_r) \) and

\[
D_r = \begin{bmatrix}
\frac{\partial \mu_r}{\partial b'} & 0 \\
0 & \frac{\partial v_r}{\partial a'}
\end{bmatrix}.
\]

The first three elements of the lower-right-hand deviation matrix, \( D_r \), are given in table 2. The components of \( \partial v_r/\partial \sigma^2 \) are given by the coefficient of relationship, while \( \partial v_r/\partial \sigma^2 \) is 1 for a diagonal element of \( V_r \) and is 0 otherwise. Because the data are expected to be approxi-
Table 2

<table>
<thead>
<tr>
<th>Relative Pair</th>
<th>$\partial \langle X_i, X_j \rangle / \partial \theta$</th>
<th>$\partial \langle X_i, X_j \rangle / \partial \sigma^2_{\theta}$</th>
<th>$\partial \langle X_i, X_j \rangle / \partial \sigma^2_{\theta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sibs</td>
<td>$2(1-20)(1-2n_x)\sigma^2_{\theta}$</td>
<td>$28(1-6)+(1-20)^2\pi_x-\frac{1}{4}$</td>
<td>$28(1-6)+(1-20)^2\pi_x$</td>
</tr>
<tr>
<td>Half-sibs</td>
<td>$(1-20)\pi_x\sigma^2_{\theta}$</td>
<td>$0$</td>
<td>$6(1-6)+(1-20)^2\pi_x$</td>
</tr>
<tr>
<td>Avuncular</td>
<td>$\frac{1}{4}(1-20)(5-6(1-20)(1-2n_x)\sigma^2_{\theta}$</td>
<td>$0$</td>
<td>$\frac{1}{4}(1-20)^2\pi_x + (1-20)^2(1-6)\pi_x$</td>
</tr>
<tr>
<td>Grandparental</td>
<td>$\frac{1}{8}(1-4n_x)\sigma^2_{\theta}$</td>
<td>$0$</td>
<td>$\frac{1}{8}(1-20)\pi_x$</td>
</tr>
<tr>
<td>Cousins</td>
<td>$\frac{1}{4}(1-20)(4-76+40\pi_x)\sigma^2_{\theta}$</td>
<td>$0$</td>
<td>$\frac{1}{4}(1-20)^2\pi_x + (1-20)^2(1-6)\pi_x$</td>
</tr>
</tbody>
</table>

mately normally distributed, the Gaussian working model (Prentice and Zhao 1991) can be applied to efficiently estimate the covariance parameters, and this yields

$$W_r = \begin{bmatrix} V_r & 0 \\ 0 & \omega_{rr} \end{bmatrix}$$

for the $i,j$th relative pair

$$\omega_{rkr} = \left\{ \begin{array}{ll} 2\sigma^2_{\theta} & \text{for the } i,j \text{th relative pair} \\ \sigma_{\theta} \sigma_{rr} + \sigma_{rr} \sigma_{rr} & \text{for the } i,j \text{th and } l,m \text{th pairs.} \end{array} \right.$$  

The covariances are given by equation (4), Table 1, and the coefficient of relationship. By this approach, estimation proceeds by an iterative procedure in which the terms affecting the kurtosis are reestimated and would theoretically be more efficient than not estimating these higher-order terms. However, this approach would require inverting $n_x + n_y$ squared matrices. Alternatively, one might force the higher moments to be independent.

In this scheme, hypotheses can be tested by robust score tests (Liang and Zeger 1986) to allow for possible deviation from multivariate normality. Provided the fixed and random effects are independent, the robust variance of $\hat{a} - a$ is given by

$$\left( \sum_{r} \frac{\partial \bar{v}_r}{\partial a} \omega^{-1} \frac{\partial \bar{v}_r}{\partial a} \right)^{-1} \times \left( \sum_{r} \frac{\partial \bar{v}_r}{\partial a} \omega^{-1}(s_r-v_r)(s_r-v_r) \omega^{-1} \frac{\partial \bar{v}_r}{\partial a} \right) \times \left( \sum_{r} \frac{\partial \bar{v}_r}{\partial a} \omega^{-1} \frac{\partial \bar{v}_r}{\partial a} \right)^{-1}.$$  

The hypothesis that a particular element of $\alpha$ is 0 could be tested by dividing the parameter by the square root of the appropriate diagonal element and then comparing it with a $t$ distribution. Although this approach shows promise in allowing for the non-normality of the data imposed by segregation of a major gene, evaluating the size of the tests will require extensive simulation studies. Olson and Wijman (1993) have shown that tests constructed by this approach had an excess size in small samples, when the higher moments were jointly estimated.

Simulation Studies

I have performed simulation studies to verify that the tests constructed using maximum-likelihood methods were not inappropriately large, because the maximum-likelihood methods applied here assume multivariate normality, even though segregation of a major gene in the data induces platykurtosis. I have also used simulation studies both to assess how some deviations from the assumed model affected the size and power of the tests to detect linkage and to evaluate the estimates.

In the simulation studies, the trait data were assumed to arise from a locus with two alleles of equal frequency. Trait alleles were randomly assigned to parents. Trait values at this quantitative locus were assigned following a normal distribution, as given by the assumed nongenetic variance with mean values conditional on trait genotypes as given by the assumed genotypic effects. No covariate effects were included. The marker locus was assumed to have four alleles of equal frequency that were either randomly assigned to parents or assigned so that all matings were of the type 1-2 X 3-4; that is, each parental allele was unique. Trait alleles were randomly segregated to offspring, and the marker alleles were assigned on the basis of the trait alleles, allowing for possible recombination events when a recombination fraction greater than 0 was assumed. For each analysis, 40 nuclear families consisting of six sibs and their parents were analyzed, and the simulation study comprised 100 replicates. Simulation parameters given in tables 3 and 4 yield $\sigma_x = 2$ for
Table 3

Power and Significance of Simulation Studies

<table>
<thead>
<tr>
<th>Simulation Experiment</th>
<th>Generating Model Trait Means</th>
<th>% of Tests Rejected at Nominal Level of</th>
<th>Median Parameter Estimates Under</th>
<th>Estimated Genetic Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu_{AA} )</td>
<td>( \mu_{AB} )</td>
<td>( \mu_{BB} )</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>1</td>
<td>2 4 6 0 .5</td>
<td>8 7 2</td>
<td>0.00 1.74 .00</td>
<td>1.80 0.00</td>
</tr>
<tr>
<td>2</td>
<td>2 4 6 0 .0</td>
<td>100 100 100</td>
<td>1.98 0.00</td>
<td>1.64 1.00</td>
</tr>
<tr>
<td>3</td>
<td>2 4 6 1 .5</td>
<td>3 2 0</td>
<td>.03 1.68 1.07</td>
<td>1.83 1.07</td>
</tr>
<tr>
<td>4</td>
<td>2 4 6 1 .0</td>
<td>100 100 100</td>
<td>1.69 .00</td>
<td>1.88 1.03</td>
</tr>
<tr>
<td>5</td>
<td>2 4 6 1 .5</td>
<td>5 3 1</td>
<td>.00 1.45</td>
<td>1.37 1.90</td>
</tr>
<tr>
<td>6</td>
<td>2 4 6 2 .0</td>
<td>94 92 81</td>
<td>1.46 .00</td>
<td>1.50 1.99</td>
</tr>
<tr>
<td>7</td>
<td>2 4 6 2 .5</td>
<td>7 4 0</td>
<td>.00 1.67</td>
<td>1.84 2.17</td>
</tr>
<tr>
<td>8</td>
<td>2 4 6 2 .0</td>
<td>96 93 86</td>
<td>1.96 .00</td>
<td>1.71 2.39</td>
</tr>
<tr>
<td>9</td>
<td>2 4 6 2 .5</td>
<td>2 1 0</td>
<td>.00 .29</td>
<td>.38 2.51</td>
</tr>
<tr>
<td>10</td>
<td>2 4 6 2 .0</td>
<td>41 29 11</td>
<td>.30 .00</td>
<td>.29 2.95</td>
</tr>
<tr>
<td>11</td>
<td>2 4 6 2 .5</td>
<td>5 4 1</td>
<td>.04 8.01</td>
<td>8.27 8.00</td>
</tr>
<tr>
<td>12</td>
<td>2 4 6 2 .0</td>
<td>96 96 89</td>
<td>4.84 3.90</td>
<td>8.55 .00</td>
</tr>
<tr>
<td>13</td>
<td>2 4 6 0 .1</td>
<td>100 100 100</td>
<td>1.50 .00</td>
<td>.33 1.95</td>
</tr>
<tr>
<td>14</td>
<td>2 4 6 0 .0</td>
<td>100 100 100</td>
<td>1.99 .00</td>
<td>1.64 .10</td>
</tr>
<tr>
<td>15</td>
<td>2 4 6 1 .0</td>
<td>100 100 100</td>
<td>1.74 .00</td>
<td>.93 1.88 1.03</td>
</tr>
</tbody>
</table>

* Parental alleles are unique.

Simulations 1–6 and 13–15 of table 3 and for all simulations of table 4. In simulations 7–8, \( \sigma^2_a = 2.0 \) and \( \sigma^2_b = 1.0 \), while for simulations 9–10, \( \sigma^2_a = 0 \) and \( \sigma^2_b = 0.5 \), and in simulations 11 and 12, \( \sigma^2_a = 8.0 \) and \( \sigma^2_b = 16.0 \). The data were analyzed by capturing output from the SIBPAL program (Elston et al. 1986), which estimated identity-by-descent sharing for pairs of sibs, including parental marker information. The direct search routine of the program MAXFUN (Elston et al. 1986) was used to calculate maximum-likelihood estimates of all parameters. In the analyses shown in table 4, the genetic parameters were constrained to the moment estimator of the within-sibship covariance. The estimator used was

\[
\sum_k \left( \frac{\sum_{i<j} (X_{ik}X_{jk}) - N \bar{X}^2}{N} \right),
\]

where \( N \) is the total number of unique sib pairs in the data, and \( \bar{X} \) is the overall mean for sibs.

The results of the simulation studies are given in ta-

Table 4

Power and Significance of Simulation Studies, Total Genetic Variance Fixed to Empirical Covariance

<table>
<thead>
<tr>
<th>Simulation Experiment</th>
<th>Generating Model Trait Means</th>
<th>% of Tests Rejected at Nominal Level of</th>
<th>Median Parameter Estimates Under</th>
<th>Estimated Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu_{AA} )</td>
<td>( \mu_{AB} )</td>
<td>( \mu_{BB} )</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>1</td>
<td>2 4 6 1 .5</td>
<td>3 2 0</td>
<td>.00 1.85 .95</td>
<td>1.92 .95</td>
</tr>
<tr>
<td>2</td>
<td>2 4 6 .5</td>
<td>6 1 0</td>
<td>.00 1.86 .95</td>
<td>1.92 .95</td>
</tr>
<tr>
<td>3</td>
<td>2 4 6 1 .0</td>
<td>100 100 100</td>
<td>1.56 .04</td>
<td>1.92 .95</td>
</tr>
<tr>
<td>4</td>
<td>2 4 6 .0</td>
<td>100 100 100</td>
<td>1.57 .07</td>
<td>1.92 .95</td>
</tr>
</tbody>
</table>

* Parental alleles are unique.
ble 3. In this analysis, no dominance components were modeled. As can be seen, under the null hypothesis of no linkage, the size of the tests is slightly conservative: the size at the nominal .05 and .01 significance levels had average empirical significances of .035 and .007, respectively. There was bias in the estimates of the proportion of variance attributable to major-gene effects in all analyses that included nongenetic variability, because some of the variance was incorrectly attributed to "polygenic" factors. The median estimates for each simulation study that are provided reflect the assignment of some variance to polygenic factors. This leads to an apparent bias in the median total genetic variance, because of the non-normality of the distribution of $\sigma^2_1$, especially when $\theta = .5$ is used to generate the data. However, the median estimate of the total genetic variance (i.e., $\sigma^2_0 + \sigma^2_1$) was also underestimated in most cases. The data generated with the major gene unlinked to the marker locus generally produced a larger underestimate of the total genetic variance than was observed when the marker and trait loci were linked. In the three models with $\mu_{AA} = 2$, $\mu_{AB} = 6$, $\mu_{BB} = 6$; $\mu_{AA} = 4$, $\mu_{AB} = 6$, and $\mu_{BB} = 4$; and $\mu_{AA} = 2$, $\mu_{AB} = 6$ and $\mu_{BB} = -6$, there was a component of dominance variance. However, in each case, under linkage, there was evidence for a major locus, and, under no linkage, there were no excess false-positive results. Ensuring that all mating types were informative at the marker locus did not greatly alter the estimates of the genetic effects.

Because the genetic effects were underestimated when the trait and marker loci were unlinked, I performed simulation studies under a model in which the total genetic variance was fixed to twice the observed covariance among sibs (table 4). For these analyses, a single set of random-number seeds was used throughout, so that the total collection of families in each analysis had the same set of assigned major genotypes. Under the constraint that total empirical and estimated covariances are equal, the variance attributable to an unlinked locus appeared to be estimated with little bias. However, under linkage and the full model, which includes all variance components, slightly more of the variance appeared to be inappropriately assigned to the "polygenic" factors. Ensuring that all of the matings were fully informative for the marker locus only slightly improved the estimate of the parameters. The nominal significance of tests performed in this manner appears to be conservative.

Conclusions

The results of these studies showed that a variance-components procedure could be used to assess evidence for genetic linkage and that this procedure showed reasonable power and no excess size. Paradoxically, the estimate of the major-gene component of variance was least biased when there was no variance attributable to nongenetic sources, i.e., when the data were most platykurtotic. Generally, the maximum-likelihood estimates of the genetic sources of variability were underestimated, whether full information about identity-by-descent sharing from the marker was available. Constraining the genetic parameters to yield estimates for which the total covariance among sibs equaled the observed covariance eliminated this bias and did not increase the size of the tests or decrease their power.

The approach, provided here, for genetic linkage can be applied to any collection of randomly selected outbred pedigrees. Simulation studies were not practicable for relative pairs other than sibs because no program was available to obtain their identity-by-descent sharing. In analysis of data, identity-by-descent sharing can be obtained (Amos et al. 1990), but computer algorithms are needed for simulation studies. The approach developed here will directly accommodate information from all relatives in a pedigree and, moreover, could be used to estimate $\theta$ in that case.

Unlike the Haseman-Elston approach, the model that I have described here is general enough to also accommodate gene-environment interactions and other processes that might induce correlation between genetic and environmental factors, such as assortative mating, but the forms of the estimators would need to be modified, since the first and second moments would no longer be independent. Similarly, the estimators may require modification if selection occurs through probands and if polygenic factors are present, since this selection process affects both the means and variances of relatives of the probands. Although the method can be adapted to estimate parameters of gene-environment interaction, specialized sampling strategies are usually required in population studies to increase the precision of the estimates of interaction terms. The approach developed here provides estimators of the components of variability attributable to polygenic, environmental, and linked monogenic factors. The variance-components approach detected genetic linkage under genetic models that would probably not be considered in many full maximum-likelihood genetic linkage studies and may therefore be considered less model dependent than full maximum-likelihood methods. This variance-components approach required iterative estimation procedures and consequently required much more computation than the Haseman-Elston approach, though
considerably less than fully model-dependent maximum-likelihood segregation and linkage methods (Amos and Laing, in press).

The approach presented here is quite similar to that of Goldgar (1990); however, the present model was developed with reference to a particular major genetic locus, while Goldgar's approach is appropriate for polygenic factors on chromosomal regions and has not been developed for relatives other than sibs. The current approach could be extended to include multiple markers when an algorithm to estimate, from multiple markers, the identity-by-descent sharing at a particular genetic location is developed. Further research is needed to explore the efficacy of various procedures for estimating the variance components and to compare their power to detect genetic linkage with that of other approaches, such as full maximum-likelihood approaches and the Haseman-Elston approach. Constraining the maximum-likelihood estimates to the empirical covariance to reduce dependence on modeling assumptions has been advocated by others (Carroll and Ruppert 1982).

Finally, although I developed the forms of estimators for use in generalized estimating-equation approaches, the properties of these tests should be studied by simulation studies, and test procedures based on quasi-likelihood procedures are likely to converge slowly to asymptotic distributions. Thus, maximum-likelihood procedures might yield relatively poor estimators for the variance components and yet provide a more conservative test for genetic linkage. Quasi-likelihood approaches allow for non-normality in the data and should provide more efficient estimates of the parameters, but they are likely to yield anticonservative hypothesis tests unless large data sets are available.

Acknowledgments

I thank Dakai Zhu for assistance in preparing this document, Sherri Bale, Maureen Goode, and the anonymous reviewers provided invaluable critical comments.

References

Amos CI, Laing AE. Comparison of univariate and multivariate tests for genetic linkage. Genet Epidemiol (in press)
Fisher RT (1918) Correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 399–433
MacCluer JW (1989) Statistical approaches to identifying