Genetic Epidemiology 26: 265–272 (2004)

Quantitative Trait Linkage Analysis by Generalized Estimating Equations: Unification of Variance Components and Haseman-Elston Regression

Wei-Min Chen,[1] Karl W. Broman,[1*] and Kung-Yee Liang[1,2]

[1]*Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland*
[2]*Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, Taiwan*

Two of the major approaches for linkage analysis with quantitative traits in humans include variance components and Haseman-Elston regression. Previously, these were viewed as quite separate methods. We describe a general model, fit by use of generalized estimating equations (GEE), for which the variance components and Haseman-Elston methods (including many of the extensions to the original Haseman-Elston method) are special cases, corresponding to different choices for a working covariance matrix. We also show that the regression-based test of Sham et al. ([2002] Am. J. Hum. Genet. 71:238–253) is equivalent to a robust score statistic derived from our GEE approach. These results have several important implications. First, this work provides new insight regarding the connection between these methods. Second, asymptotic approximations for power and sample size allow clear comparisons regarding the relative efficiency of the different methods. Third, our general framework suggests important extensions to the Haseman-Elston approach which make more complete use of the data in extended pedigrees and allow a natural incorporation of environmental and other covariates. © 2004 Wiley-Liss, Inc.

Grant sponsor: National Institutes of Health; Grant number: GM49909.
*Correspondence to: Karl W. Broman, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205.
E-mail: kbroman@jhsph.edu
Received 30 October 2003; Accepted 1 December 2003
Published online 25 February 2004 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.10315

INTRODUCTION

Many important human disease-related phenotypes (e.g., blood pressure) are quantitative in nature. There are a plethora of approaches for linkage analysis of quantitative traits in human data, but until recently, there was a dearth of understanding of the advantages and disadvantages of the different approaches; two recent reviews [Feingold, 2001, 2002] were especially valuable in improving this understanding.

Two of the most commonly used approaches for quantitative trait linkage analysis are Haseman-Elston regression [Haseman and Elston, 1972] and the use of variance components models [Amos, 1994; Almasy and Blangero, 1998]. Previously, these approaches were viewed as completely separate methods. In this paper, we describe a general method for quantitative trait linkage analysis that makes use of generalized estimating equations (GEE) [Liang and Zeger, 1986], for which the variance components method and Haseman-Elston regression (including many of its extensions) are special cases. This work has several important implications: it provides new insights into the relationship between these methods, it leads to asymptotic sample-size approximations that allow clear comparisons between methods, and it suggests important extensions to Haseman-Elston regression, both for its application in general pedigrees and for the incorporation of environmental and other covariates.

SIBLING PAIRS

We first illustrate our general approach in the special case of randomly ascertained sibling pairs with known population mean phenotype (assumed, without loss of generality, to be 0), under the assumption that there is a single putative quantitative trait locus (QTL) with no dominance effect. Let y_{k1}, y_{k2} denote the phenotypes for the kth sibling pair, with $\boldsymbol{y}_k = (y_{k1}, y_{k2})'$. Let π_k denote, for the kth pair, the proportion of alleles shared identical by descent (IBD) at a putative QTL. Let M_k denote the available multipoint marker data for the pair, and let $\hat{\pi}_k = \mathrm{E}(\pi_k|M_k)$, the expected proportion of alleles

shared IBD given the marker data. Let σ_a^2 denote the additive variance due to the putative QTL, let σ^2 denote the overall phenotypic variance, and let ρ denote the correlation between siblings' phenotypes.

In the variance components approach to quantitative trait linkage analysis [Amos, 1994; Almasy and Blangero, 1998], the phenotypes for a sibling pair, conditional on the marker data, are assumed to follow a bivariate normal distribution with the covariance matrix for the kth pair being the following:

$$\Omega_k = \begin{pmatrix} \Omega_{k1} & \Omega_{k2} \\ \Omega_{k2} & \Omega_{k1} \end{pmatrix} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 + \sigma_a^2(\hat{\pi}_k - \frac{1}{2}) \\ \rho\sigma^2 + \sigma_a^2(\hat{\pi}_k - \frac{1}{2}) & \sigma^2 \end{pmatrix}. \quad (1)$$

The log likelihood function for this model is $l(\sigma_a^2, \sigma^2, \rho) = -(1/2)\sum_k \{ln|\Omega_k| + \boldsymbol{y}_k\Omega_k^{-1}\boldsymbol{y}_k\}$. The maximum likelihood estimates (MLEs) of the parameters, σ_a^2, ρ, and σ^2 are the values for which this function achieves its maximum, and are obtained as the solutions of the score equations:

$$0 = \frac{\partial l}{\partial \sigma_a^2} = \sum_k \left(\hat{\pi}_k - \frac{1}{2}\right)\left\{\frac{(y_{k1}+y_{k2})^2}{4(\Omega_{k1}+\Omega_{k2})^2} - \frac{(y_{k1}-y_{k2})^2}{4(\Omega_{k1}-\Omega_{k2})^2} + \frac{\Omega_{k2}}{\Omega_{k1}^2-\Omega_{k2}^2}\right\}. \quad (2)$$

$$0 = \frac{\partial l}{\partial \rho} = \sigma^2 \sum_k \left\{\frac{(y_{k1}+y_{k2})^2}{4(\Omega_{k1}+\Omega_{k2})^2} - \frac{(y_{k1}-y_{k2})^2}{4(\Omega_{k1}-\Omega_{k2})^2} + \frac{\Omega_{k2}}{\Omega_{k1}^2-\Omega_{k2}^2}\right\}. \quad (3)$$

$$0 = \frac{\partial l}{\partial \sigma^2} = \rho \sum_k \left\{\frac{(y_{k1}+y_{k2})^2}{4(\Omega_{k1}+\Omega_{k2})^2} - \frac{(y_{k1}-y_{k2})^2}{4(\Omega_{k1}-\Omega_{k2})^2} + \frac{\Omega_{k2}}{\Omega_{k1}^2-\Omega_{k2}^2}\right\} + \sum_k \frac{(y_{k1}+y_{k2})^2}{2(\Omega_{k1}+\Omega_{k2})^2} - \frac{(y_{k1}-y_{k2})^2}{2(\Omega_{k1}-\Omega_{k2})^2} - \frac{2\Omega_{k1}}{\Omega_{k1}^2-\Omega_{k2}^2}\right\}. \quad (4)$$

A more general method, making use of generalized estimating equations (GEE) [Liang and Zeger, 1986, Prentice and Zhao, 1991], can lead to this same set of equations. GEE were developed for the analysis of longitudinal data, where there are multiple measurements with known correlation structure, but for which the correlations may depend on a set of parameters that are to be estimated. Consider as the outcome for the kth sibling pair $z_k = (y_{k1}^2, y_{k2}^2, y_{k1}y_{k2})'$. With our simplifying assumption that the population phenotype mean is 0, z_k has expected value, given the observed marker data, $\mathrm{E}(z_k|M_k) = (\Omega_{k1}, \Omega_{k1}, \Omega_{k2})'$. (Recall, from Equation (1), that $\Omega_{k1} = \sigma^2$ and $\Omega_{k2} = \rho\sigma^2 + \sigma_a^2(\hat{\pi}_k - 1/2)$.)

GEE make use of a working covariance matrix, W_k, which is a set of presumed variances and covariances for the elements of z_k, and which may include unknown parameters that are to be estimated. Having specified W_k, which can be any symmetric, positive definite matrix, the GEE estimators of the parameters, σ_a^2, ρ, and σ^2, are obtained by solving the equation

$$\sum_k D_k W_k^{-1} S_k = 0 \quad (5)$$

where $S_k = z_k - \mathrm{E}(z_k|M_k)$ and D_k is a matrix whose columns consist of the derivatives of the vector $\mathrm{E}(z_k|M_k)$ with respect to each parameter, so that, in the case under consideration, and with the parameters ordered σ_a^2, ρ, σ^2,

$$D_k = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ \hat{\pi}_k - 1/2 & \sigma^2 & \rho \end{pmatrix}.$$

Different choices of working covariance matrix, W_k, lead to different parameter estimates. In particular, if the working covariance matrix has the form

$$W_k^{VC} = \begin{pmatrix} 2\Omega_{k1}^2 & 2\Omega_{k2}^2 & 2\Omega_{k1}\Omega_{k2} \\ 2\Omega_{k2}^2 & 2\Omega_{k1}^2 & 2\Omega_{k1}\Omega_{k2} \\ 2\Omega_{k1}\Omega_{k2} & 2\Omega_{k1}\Omega_{k2} & \Omega_{k1}^2 + \Omega_{k2}^2 \end{pmatrix}$$

then, through relatively straightforward algebra (e.g., by use of the computer program MATHEMATICA), one may show that Equation (5) corresponds exactly to the score equations for the variance components approach, (2)–(4). Thus the variance components method is a special case of this more general GEE method.

Note that the usual estimated standard errors (SEs) for the variance components method may be obtained via the matrix $(\sum_k D_k'(W_k^{VC})^{-1}D_k)^{-1}$.

Alternatively, we recommend the use of the more robust "sandwich" estimates, commonly used for the GEE method,

$$\left(\sum_k D_k' W_k^{-1} D_k\right)^{-1} \left\{\sum_k \left(D_k' W_k^{-1} S_k\right)\left(D_k' W_k^{-1} S_k\right)'\right\} \times \left(\sum_k D_k' W_k^{-1} D_k\right)^{-1}. \quad (6)$$

In the original Haseman-Elston method [Haseman and Elston, 1972], one uses linear regression of the squared difference between the siblings' phenotypes, $(y_{k1} - y_{k2})^2$, on the expected proportion of alleles shared IBD at the putative QTL, $\hat{\pi}_k$. The slope obtained by ordinary least squares (OLS) is an estimate of $-2\sigma_a^2$. (Note that one cannot obtain separate estimates of ρ and σ^2 by this approach, but only of the combination $(1-\rho)\sigma^2$.) Consider the following working covariance matrix

$$W^{HE} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 1/2 & 3/2 \end{pmatrix}.$$

The insertion of W^{HE} as the working covariance matrix in Equation (5) leads to the following:

$$0 = \sum_k \left(\hat{\pi}_k - \frac{1}{2}\right) \times \left\{-\frac{(y_{k1} - y_{k2})^2}{2} + \Omega_{k1} - \Omega_{k2}\right\} \quad (7)$$

$$0 = \sum_k \sigma^2 \left\{-\frac{(y_{k1} - y_{k2})^2}{2} + \Omega_{k1} - \Omega_{k2}\right\} \quad (8)$$

$$0 = \sum_k \left\{(3-\rho)\left(\frac{y_{k1}^2 + y_{k2}^2}{2} - \Omega_{k1}\right) - (1-\rho)(y_{k1}y_{k2} - \Omega_{k2})\right\} \quad (9)$$

Equation (9) turns out to be redundant, and the solution of Equations (7) and (8) for σ_a^2 and $(1-\rho)\sigma^2$ give estimates that are identical to those derived from the original Haseman-Elston method. Thus, Haseman-Elston is a special case of our general GEE approach, corresponding to the use of the working covariance matrix W^{HE}.

The usual estimated SE used with Haseman-Elston regression is that from ordinary least squares (OLS), based on the assumption of constant variance, which is correct under the null hypothesis of no linkage, but is generally not correct under the alternative hypothesis that the site under test is linked to a QTL. The estimated SE from our GEE method, based on the sandwich estimate of the variance matrix, does not rely on the constant variance assumption and provides a consistent estimate of the SE even in the case of linkage.

Wright [1997] pointed out that further information may be obtained by considering the squared sum of the siblings' quantitative phenotypes, in addition to the squared difference. Several extensions to the original Haseman-Elston method take advantage of this observation. In the Haseman-Elston revisited method [Elston et al., 2000], the product of the siblings' phenotypes, $y_{k1}y_{k2}$, is regressed on the expected proportion of alleles shared IBD at the putative QTL, $\hat{\pi}_k$. This approach is also a special case of our general GEE method, corresponding to use of the identity matrix as the working covariance matrix.

A further extension of the original Haseman-Elston method is the combined Haseman-Elston regression method (denoted HE-COM) of Sham and Purcell [2001]. In this method, ρ and σ^2 are assumed known, and one regresses $(y_{k1} + y_{k2})^2/(1+\rho)^2 - (y_{k1} - y_{k2})^2/(1-\rho)^2$ on $\hat{\pi}_k$ to obtain an estimate of σ_a^2. Consider the following working covariance matrix:

$$W^{COM} = \begin{pmatrix} \frac{1+\rho^2}{(1-\rho^2)^2\sigma^4} & -\frac{1+\rho^2}{(1-\rho^2)^2\sigma^4} & 0 \\ -\frac{1+\rho^2}{(1-\rho^2)^2\sigma^4} & \frac{(1+\rho^2)(1+4(1+\rho^2)\sigma^4)}{(1-\rho^2)^2\sigma^4} & \frac{4\rho(1+\rho^2)}{(1-\rho^2)^2} \\ 0 & \frac{4\rho(1+\rho^2)}{(1-\rho^2)^2} & \frac{(1+\rho^2)^2}{(1-\rho^2)^2} \end{pmatrix}.$$

Inserting the working covariance matrix W^{COM} into Equation (5) (though here we take only the first column of the matrix D_k, as only the parameter σ_a^2 remains to be estimated), one can show that this approach is also a special case of our general GEE method.

Thus, for the case of randomly ascertained sibling pairs, and with the assumption that the population phenotype mean is known (made in order to simplify the algebraic expressions), we have shown that the variance components method for quantitative trait linkage analysis, as well as the original Haseman-Elston, Haseman-Elston revisited, and HE-COM methods, are all special cases of a general GEE method.

GENERAL PEDIGREES

While we focused above on the case of sibling pairs, the results may be seen to apply more generally. Consider a set of general pedigrees with no inbreeding, and let y_{ki} denote the quantitative phenotype for the ith individual in the kth pedigree. Let Φ_{kij} and Δ_{kij} denote the kinship and fraternity coefficients, respectively, for individuals i and j in pedigree k, and let $\hat{\pi}_{kij}$ and $\hat{\kappa}_{kij}$ denote their expected proportion of alleles shared IBD and the probability that they share two alleles IBD, respectively, at a putative QTL, given multipoint marker data. Let σ_a^2 and σ_d^2 denote the additive and dominance variance, respectively, due to the putative QTL, and let σ_{pa}^2, σ_{pd}^2, and σ_e^2 denote the additive polygenic variance, dominance polygenic variance, and residual environmental variance, respectively. (Note that, for the sibling pairs case considered above, we used a different but equivalent parameterization: we assumed that $\sigma_d^2 = 0$ and considered $\sigma^2 = \sigma_a^2 + \sigma_{pa}^2 + \sigma_{pd}^2 + \sigma_e^2$ and $\rho = (\sigma_a^2/2 + \sigma_{pa}^2/2 + \sigma_{pd}^2/4)/\sigma^2$.)

Consider a set of p covariates (including an intercept term), and assume that $\mathrm{E}(y_{ki}) = \mathrm{E}(y_{ki}|M_{ki}) = \boldsymbol{x}_{ki}'\beta$. The covariance of the phenotypes for individuals i and j in pedigree k, given the available marker data, is

$$\Omega_{kij} = \begin{cases} \sigma_a^2 + \sigma_d^2 + \sigma_{pa}^2 + \sigma_{pd}^2 + \sigma_e^2 & \text{if } i = j \\ \hat{\pi}_{kij}\sigma_a^2 + \hat{\kappa}_{kij}\sigma_d^2 + 2\Phi_{kij}\sigma_{pa}^2 + \Delta_{kij}\sigma_{pd}^2 & \text{if } i \neq j \end{cases} \tag{10}$$

For mathematical convenience, we consider as the outcome for the kth pedigree $\boldsymbol{z}_k = [y_{ki}, (y_{ki} - \boldsymbol{x}_{ki}'\beta)^2, (y_{ki} - \boldsymbol{x}'_{ki}\beta)(y_{kj} - \boldsymbol{x}_{kj}'\beta)]'$, a vector of length $m_k = 2n_k + n_k(n_k - 1)/2$, where n_k is the number of phenotyped individuals in pedigree k. (There are a variety of other equivalent formulations, but this leads to somewhat simpler algebraic expressions.) Note that $\mathrm{E}(\boldsymbol{z}_k|M_k) = (\boldsymbol{x}_{ki}'\beta, \Omega_{kii}, \Omega_{kij})'$.

With our GEE method, the $p+5$ parameters (σ_a^2, σ_d^2, σ_{pa}^2, σ_{pd}^2, σ_e^2, and β) are estimated as the solutions to the same Equation (5) for some choice of working covariance matrix W_k, and again with $S_k = \boldsymbol{z}_k - \mathrm{E}(\boldsymbol{z}_k|M_k)$ (a vector of length m_k) and D_k a matrix (of dimension $m_k \times (\mathrm{p}+5)$) whose columns consist of the derivatives of the vector $\mathrm{E}(\boldsymbol{z}_k|M_k)$ with respect to each parameter (in the order referred to above), as follows:

$$D_k = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & X_k \\ 1 & 1 & 1 & 1 & 1 & 0 \\ [\hat{\pi}_{kij}] & [\hat{\kappa}_{kij}] & [2\Phi_{kij}] & [\Delta_{kij}] & 0 & 0 \end{pmatrix}.$$

Again, different choices for the working covariance matrix, W_k, lead to different estimates, and robust SEs for the GEE estimates can again be obtained via Equation (6).

In the variance components approach for quantitative trait linkage analysis in general pedigrees [Almasy and Blangero, 1998], the phenotypes $\boldsymbol{y}_k$ are assumed to follow a multivariate normal distribution with mean $X_k\beta$ and covariance matrix as in Equation (10), and the parameters are estimated by maximum likelihood. Through relatively straightforward but tedious algebra, it can be shown that the MLEs under the normal model correspond to the estimates from our general GEE method, for the case that the working covariance matrix is the following:

$$W_k^{VC} = \begin{pmatrix} \Omega_k & 0 & 0 \\ 0 & A_k & B_k \\ 0 & B_k' & C_k \end{pmatrix}.$$

Here A_k is a matrix of dimension $n_k \times n_k$ with $A_{kij} = 2\Omega_{kij}^2$. B_k is a matrix of dimension $n_k \times n_k(n_k-1)/2$ whose columns correspond to pairs of individuals; let $(s{:}t)$ denote the column corresponding to the pair (s, t) with $s < t$. Then the value in the ith row and $(s{:}t)$th column of B_k is $2\Omega_{kis}\Omega_{kit}$, the covariance, given the marker data, of y_{ki}^2 and $y_{ks}y_{kt}$ under the assumption of multivariate normality. Finally, C_k is a square, symmetric matrix with $n_k(n_k-1)/2$ rows and columns; the value in the $[(i{:}j), (s{:}t)]$ position is $\Omega_{kis}\Omega_{kjt} + \Omega_{kit}\Omega_{kjs}$.

It should be noted that Amos [1994] and Amos et al. [1996] applied GEE for quantitative trait linkage analysis, with the working covariance matrix, W_k^{VC}, though it was not recognized that this approach is identical to maximum likelihood under a normal model.

Olson and Wijsman [1993] extended the original Haseman-Elston method for use with general pedigrees, considering the squared phenotype differences for all relative pairs, and using a GEE approach with a working covariance matrix denoted here as V_k^{HE}. This can also be shown to be a special case of our general GEE method, with working covariance matrix

$$W_k^{HE} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & \frac{1}{2}E_k \\ 0 & \frac{1}{2}E_k' & \frac{1}{4}(V_k^{HE} + E_k'E_k) \end{pmatrix}$$

where E_k is a matrix of dimension $n_k \times n_k(n_k-1)/2$ whose j, $(s{:}t)$ element is 1 if $j{=}s$ or $j{=}t$, and is 0 otherwise.

DISCUSSION

We have described a general method, making use of generalized estimating equations (GEE), for quantitative trait linkage analysis in human pedigrees, which unifies the variance components and Haseman-Elston methods, as each is a special case of our general method, corresponding to different choices for the working covariance matrix. Our GEE method is similar to, but more general than, the GEE method recently described by Shete et al. [2003]. They focused on sibships, considered the squared differences and squared sums of all sibling pairs' phenotypes, and used a particular working covariance matrix.

Our GEE method generalizes and unifies the variance components and Haseman-Elston methods in the sense that the parameter estimates obtained as solutions to the GEE are identical to the MLEs for the variance components method if W^{VC} is used as the working covariance matrix, or identical to the OLS estimates for Haseman-Elston regression if W^{HE} is used as the working covariance matrix. However, the usual test statistics for linkage for the variance components and Haseman-Elston methods do not follow immediately from the GEE method.

In variance components, one typically uses the likelihood ratio test statistic, which requires that one consider directly the normal likelihood. In Haseman-Elston regression, one typically uses a Wald statistic based on the SE from ordinary least squares. Alternatively, one may use a score statistic derived from the normal likelihood, such as the robust score statistic of Wang and Huang [2002], developed particularly for sibships. While GEE method we have described does not lead directly to any of these test statistics, it does provide the parameter estimates that are the basis of any test statistic, and so any such statistic may be calculated immediately using the results of GEE. We are currently investigating the relative performance, in terms of power and robustness, of a variety of such test statistics in the case of sibships and larger pedigrees.

Sham et al. [2002] described a new method for quantitative trait linkage analysis in human pedigrees, in which the IBD status for all relative pairs is regressed on the squared differences and squared sums of the pairs' phenotypes. The method was implemented in the software MERLIN [Abecasis et al., 2002], and was shown to have power similar to the variance components approach but to be robust to departures from normality. It is intriguing to note that this method corresponds exactly to a robust score statistic that may be derived from our GEE method with the Gaussian working covariance matrix, W^{VC}. The details are deferred to the Appendix. We implemented this score test in our own software and confirmed the mathematical result: with simulated data, the test statistic was identical to the results of the software MERLIN-REGRESS.

This work has several implications, the most important of which is the new insight that it provides on the connection between the Haseman-Elston and variance components methods: choosing between these approaches is equivalent to choosing a working covariance matrix for the GEE method. In the case of multivariate normality, the variance components method will have improved power over Haseman-Elston regression, as it is based on the correct covariance matrix with no additional parameters [Liang et al., 1992]. In the absence of normality, the use of the likelihood ratio statistic with the variance components method can give an inflated type I error rate [e.g., Allison et al., 1999]. The use of GEE with robust SEs (i.e., based on the sandwich estimator) will control the type I error rate; as a special case, Haseman-Elston regression is robust. As the working covariance matrix for the variance components method, W^{VC}, will still likely be closer to the truth than that of Haseman-Elston regression, W^{HE}, even when the normal model is not correct, one may use GEE with the working covariance matrix W^{VC} to obtain a method that is as robust as Haseman-Elston regression in terms of type I error, but has higher power [Liang et al., 1992].

In addition, our general GEE method provides an approach for extending the Haseman-Elston method to general pedigrees that makes more full use of the available data than the method of Olson and Wijsman [1993], and allows the incorporation of environmental covariates. A careful assessment of the advantages and disadvantages of different choices for a working covariance matrix deserves further exploration.

Finally, the unification of a variety of quantitative trait linkage analysis methods within a single

general framework enables a more simple comparison of the relative performance of the methods. As an example, we consider the case of n sibling pairs (though note that these results may be easily extended for the case of general pedigrees). We assume that the siblings' phenotypes approximately follow a bivariate normal distribution. In this case, the Wald test statistics, with SEs from the sandwich estimator of the variance matrix, for the four methods considered above each follow, approximately, a noncentral χ^2 distribution with one degree of freedom, with noncentrality parameter (NCP) according to the following formula:

$$\mathrm{NCP} = \frac{\sigma_a^4 \left(\sum_k D_k' W_k^{-1} D_k\right)^2}{\sum_k D_k' W_k^{-1} W_k^{VC} W_k^{-1} D_k}$$

where W_k^{VC} is the working covariance matrix for the variance components method, which is the true covariance matrix under the assumption of bivariate normality. For the case of sibling pairs, algebraic expressions for the NCP may be obtained; they are displayed in Table I. Note that in the case that the QTL under study explains a small proportion of the total genetic effect (i.e., $\sigma_a^2/\sigma^2 \ll 2\rho$), these formulas reduce to the approximate formulas of Sham and Purcell [2001], listed in the third column of Table I, in which case their method, HE-COM, was seen to be equivalent to the variance components method. With the more precise formulas in the middle column of Table I, the HE-COM method can be seen to have slightly lower power than the variance components method.

TABLE I. Noncentrality parameters for case of n sibling pairs, under a normality assumption

Method	Noncentrality parameter	Approximation
H-E	$\frac{n(\sigma_a^2/\sigma^2)^2}{16(1-\rho)^2+4(\sigma_a^2/\sigma^2)^2}$	$\frac{n}{16(1-\rho)^2}\left(\frac{\sigma_a^2}{\sigma^2}\right)^2$
H-E revisited	$\frac{n(\sigma_a^2/\sigma^2)^2}{8(1+\rho^2)+2(\sigma_a^2/\sigma^2)^2}$	$\frac{n}{8(1+\rho)^2}\left(\frac{\sigma_a^2}{\sigma^2}\right)^2$
HE-COM	$\frac{n(1+\rho^2)^2(\sigma_a^2/\sigma^2)^2}{8(1-\rho^2)^2(1+\rho^2)+2(1+6\rho^2+\rho^4)(\sigma_a^2/\sigma^2)^2}$	$\frac{n(1+\rho^2)}{8(1-\rho^2)^2}\left(\frac{\sigma_a^2}{\sigma^2}\right)^2$
VC	$\frac{n}{16}\left(\frac{\sigma_a^2}{\sigma^2}\right)^2\left\{\frac{1+(\rho+\sigma_a^2/2\sigma^2)^2}{\left[1-(\rho+\sigma_a^2/2\sigma^2)^2\right]^2}+\frac{1+(\rho-\sigma_a^2/2\sigma^2)^2}{\left[1-(\rho-\sigma_a^2/2\sigma^2)^2\right]^2}\right\}$	$\frac{n(1+\rho^2)}{8(1-\rho^2)^2}\left(\frac{\sigma_a^2}{\sigma^2}\right)^2$

[a]Approximation for case $\sigma_a^2/\sigma^2 \ll 2\rho$.

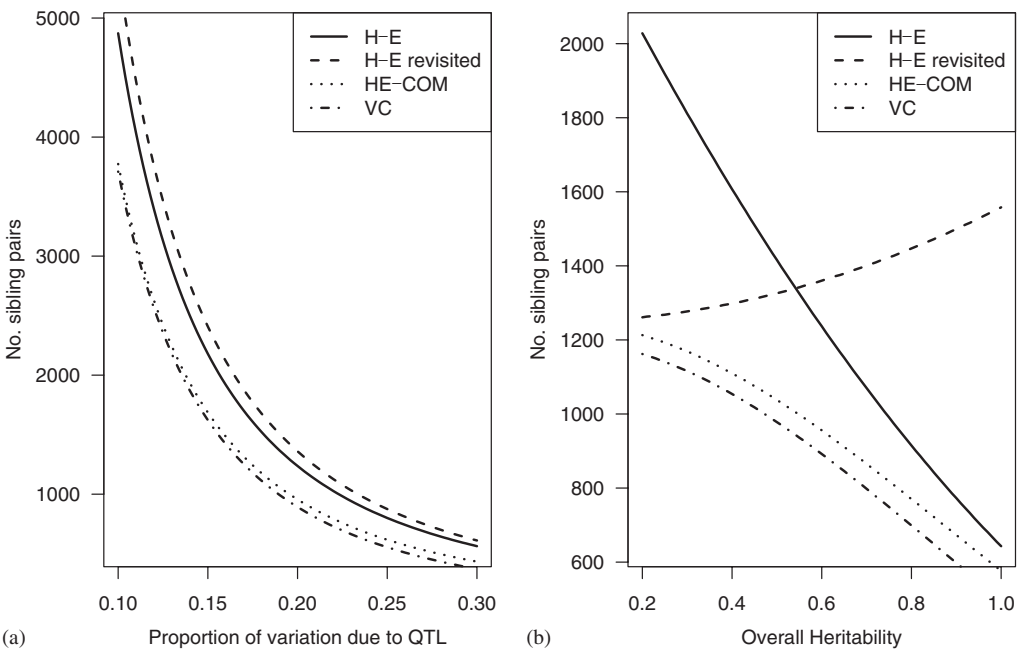

Fig. 1. Number of sibling pairs required to achieve 80% power to detect a QTL, for four different linkage analysis methods. A: Sample size as a function of proportion of phenotypic variance due to the QTL, with an overall heritability of 60%. B: Sample size as a function of overall heritability, with 20% of phenotypic variance due to the QTL.

The sample size required to achieve power $1-\beta$ with significance level α is obtained by solving the equation $\text{NCP} = (Z_\alpha - Z_{1-\beta})^2$ for the sample size, n. Figure 1 displays the number of sibling pairs required to achieve 80% power to detect a QTL. In Figure 1A, the overall heritability is taken to be 60%, and the effect of the QTL is varied. In Figure 1B, the effect of the QTL is fixed at 20%, and the overall heritability is varied. As observed previously [e.g., Allison et al., 1999], the variance components approach is seen to have the greatest power in this situation; the HE-COM method performs nearly as well.

There is a great deal of flexibility in the general GEE method, as described in this paper. It will be valuable to explore the power and robustness properties of this method with different choices for the working covariance matrix, in order to identify a quantitative trait linkage analysis procedure that is as robust as Haseman-Elston regression but maintains the power of the variance components approach.

ACKNOWLEDGEMENTS

The authors thank an anonymous reviewer for valuable suggestions for improving the manuscript.

REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. MERLIN—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101.

Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J. 1999. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet 65:531–544.

Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211.

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

Amos CI, Zhu DK, Boerwinkle E. 1996. Assessing genetic linkage and association with robust components of variance approaches. Ann Hum Genet 60:143–160.

Elston RC, Buxbaum S, Jacobs KB, Olson JM. 2000. Haseman and Elston revisited. Genet Epidemiol 19:1–17.

Feingold E. 2001. Methods for linkage analysis of quantitative trait loci in humans. Theor Popul Biol 60:167–180.

Feingold E. 2002. Regression-based quantitative-trait-locus mapping in the 21st century. Am J Hum Genet 71:217–222.

Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19.

Liang KY, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13–22.

Liang KY, Zeger SL, Qaqish B. 1992. Multivariate regression analyses for categorical data J Royal Statis Soc B 54:3–40.

Olson JM, Wijsman EM. 1993. Linkage between quantitative trait and marker loci: methods using all relative pairs. Genet Epidemiol 10:87–102.

Prentice RL, Zhao LP. 1991. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 47:825–839.

Sham PC, Purcell S. 2001. Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. Am J Hum Genet 68:1527–1532.

Sham PC, Purcell S, Cherny SS, Abecasis GR. 2002. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 71:238–253.

Shete S, Jacobs KB, Elston RC. 2003. Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: weighting sums and differences. Hum Hered 55:79–85.

Wang K, Huang J. 2002. A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. Am J Hum Genet 70:412–424.

Wright FA. 1997. The phenotypic difference discards sib-pair QTL linkage information. Am J Hum Genet 60:740–742.

APPENDIX

Sham et al. [2002] proposed a method for quantitative trait linkage analysis in which IBD status is regressed upon the squared differences and squared sums of relative-pairs' phenotypes. Here we show that this method is equivalent to a score test that may be derived from our GEE approach.

For a family with n individuals, let Y denote a vector containing the $n(n-1)/2$ squared sums and n of the squared differences of the phenotypes for all relative pairs, and let $\hat{\mathbf{\Pi}}$ denote a matrix of IBD probabilities for all pairs. Let $Y_c = Y - E(Y)$, $\hat{\mathbf{\Pi}}_c = \hat{\mathbf{\Pi}} - 2\Phi$, where Φ is a matrix of kinship coefficients, and let Σ_Y denote the covariance matrix for Y, assuming that the trait values follow a multivariate normal distribution. Further define a matrix $\Sigma_{\hat{\Pi}_c}$ with elements

$$\begin{aligned}\text{Cov}[\hat{\pi}_{ij}, \hat{\pi}_{lm}] &= \text{Cov}(E[\pi_{ij}|M], E[\pi_{lm}|M]) \\ &\approx \text{Cov}(\pi_{ij}, \pi_{lm}) - \text{Cov}(\pi_{ij}, \pi_{lm}|M) \\ &= \text{Cov}(\pi_{ij}, \pi_{lm}) - (E[\pi_{ij}\pi_{lm}|M] - \hat{\pi}_{ij}\hat{\pi}_{lm})\end{aligned}$$

where $\text{Cov}(\pi_{ij},\pi_{lm})$ can be calculated given only the pedigree structure and $E[\pi_{ij},\pi_{lm}|M]$ can be calculated based on the posterior distribution conditional on marker information M. Finally, define

$$H = \begin{pmatrix} 2I_n & 0 & -2I_n \\ 0 & 2I_{n(n-3)/2} & 0 \end{pmatrix}.$$

Then the test statistic of Sham et al. [2002] is the following:

$$T = \frac{\left(\sum \hat{\mathbf{\Pi}}_c' H \Sigma_Y^{-1} Y_c\right)^2}{\sum \left[Y_c' \Sigma_Y^{-1} H' \Sigma_{\hat{\mathbf{\Pi}}} H \Sigma_Y^{-1} Y_c \right]}. \quad (11)$$

We seek to show that statistic (11) is identical to the following score test statistic:

$$S = \frac{\left(\sum \begin{pmatrix} 0 & \hat{\mathbf{\Pi}}_c' \end{pmatrix} G_0^{-1}(z - E[z])\right)^2}{\sum \left((z - E[z])' G_0^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{\hat{\mathbf{\Pi}}} \end{pmatrix} G_0^{-1}(z - E[z]) \right)} \quad (12)$$

where z is a vector consisting of all squares and cross products of trait values, and G_0 is the covariance matrix of z, assuming that the trait values follow a multivariate normal distribution.

There exists a nonsingular matrix A such that $Y = Az$. Thus $Y_c = A(z - E[z])$ and $\Sigma_Y = AG_0A'$. By straightforward algebra, we can show the following:

$$\frac{\partial E[Y|M]}{\partial \sigma_a^2} = H' \hat{\mathbf{\Pi}}_c$$

$$\frac{\partial E[z|M]}{\partial \sigma_a^2} = \begin{pmatrix} 0 \\ \hat{\mathbf{\Pi}}_c \end{pmatrix}.$$

It follows that

$$H' \hat{\mathbf{\Pi}}_c = A \begin{pmatrix} 0 \\ \hat{\mathbf{\Pi}}_c \end{pmatrix}$$

$$H' \Sigma_{\hat{\mathbf{\Pi}}} H = A \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{\hat{\mathbf{\Pi}}} \end{pmatrix} A'.$$

Thus, the square root of the numerator of statistic (11) is

$$\begin{aligned} &\sum \hat{\mathbf{\Pi}}_c' H \Sigma_Y^{-1} Y_c \\ &= \sum \begin{pmatrix} 0 & \hat{\mathbf{\Pi}}_c' \end{pmatrix} A' (A G_0 A')^{-1} A (z - E[z]) \\ &= \sum \begin{pmatrix} 0 & \hat{\mathbf{\Pi}}_c' \end{pmatrix} G_0^{-1} (z - E[z]) \end{aligned}$$

which can be shown to correspond to the generalized estimating equations with a Gaussian working covariance matrix (equivalently, to the score function) evaluated $\sigma_a^2 = 0$.

The denominator of (11) is

$$\begin{aligned} &\sum \left(Y_c' \Sigma_Y^{-1} H' \Sigma_{\hat{\mathbf{\Pi}}} H \Sigma_Y^{-1} Y_c \right) \\ &= \sum \left((z - E[z])' G_0^{-1} A^{-1} H' \Sigma_{\hat{\mathbf{\Pi}}} H A'^{-1} G_0^{-1} (z - E[z]) \right) \\ &= \sum \left((z - E[z])' G_0^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{\hat{\mathbf{\Pi}}} \end{pmatrix} G_0^{-1} (z - E[z]) \right) \end{aligned}$$

which is a robust variance estimator for the score under the null hypothesis of no linkage. It follows that the test statistic (11) is identical to the score test statistic (12).