

Chapter 1

Regression-based Linkage Analysis Methods

Tao Wang and Robert C. Elston*

*Department of Epidemiology and Biostatistics
Case Western Reserve University, Wolstein Research Building
2103 Cornell Road, Cleveland, OH 44106-7281, USA*

Regression-based methods of model-free linkage analysis offer a valuable framework for mapping both quantitative and qualitative traits. Beginning with the method proposed by Haseman and Elston,¹ these methods have been widely used in practice because of their simplicity and robustness. Furthermore, the newer methods can utilize full information from trait values, and they are applicable to any type of pedigree data. With the availability of the denser markers and appropriate sampling, these methods give hope that they may play an important role in mapping complex genetic traits. The information yielded by such an analysis can guide and facilitate the design and result inference of further association studies.

1. INTRODUCTION

Model-free linkage methods are commonly used for mapping complex diseases because they do not require the mode of inheritance of the trait under study to be correctly specified in any detail. Among the various model-free linkage methods, regression-based methods offer a valuable framework for both quantitative and qualitative traits. The regression-based linkage analysis method was first proposed by Haseman and Elston.¹ The Haseman and Elston (HE) regression is simple, robust, and therefore widely used in practice. However, there are some limitations to the original HE regression

*Correspondence author.

method. One limitation is that the original HE regression may be less powerful compared with other full likelihood-based variance component methods because it does not make full use of all the information available in the trait values. Another major limitation is that the original HE regression is not applicable to any type of pedigree data. To overcome these limitations, in recent years there have been great efforts to extend and enhance the original HE method. The newer regression methods can fully utilize the trait values on multiple types of relative pairs, and hence can play a crucial role in the linkage analysis of complex traits.

Our discussion of regression-based methods will mostly focus on quantitative trait locus (QTL) mapping. The typical regression-based method is to regress a measure of trait similarity between the members of a relative pair on a measure of their genetic similarity, which is usually described by the number of alleles shared identical by descent (IBD) at a genomic location. The rationale behind this method is that, when a marker is not linked to a disease locus, the number of marker alleles shared IBD by a pair of relatives does not depend on the relatives' trait values; on the other hand, when the marker is linked to a disease locus, we expect a correlation between the number of alleles shared IBD at the marker locus and the measure of the relatives' trait similarity. The regression methods evaluate this correlation by examining the regression coefficient. From this point of view, critical for the power and the validity of a regression-based linkage method is how we calculate the trait similarity measure and the number of marker alleles shared IBD.

In Sec. 2, we shall discuss the measure of genetic similarity and the number of alleles shared IBD. We focus on possible pitfalls in the IBD calculation, which may result in deleterious effects for all model-free-based linkage methods. In Sec. 3, we initially focus on independent sib pairs and quantitative traits, and discuss the various regression-based methods based on different definitions of the trait similarity measure in this setup. In Sec. 4, we further introduce regression-based methods recently developed to deal with any family structure. Lastly, we briefly cover regression-based methods for affected sib pairs (ASPs) that incorporate covariates.

2. THE NUMBER (PROPORTION) OF ALLELES SHARED IBD

The various regression-based linkage methods nowadays depend on the crucial concept of the number of alleles shared identical by descent (IBD), although the original model-free linkage method was based on marker

identity in state, the marker similarity between two relatives being based solely on their similarity with respect to marker phenotype. The number of alleles shared IBD is defined as the number of alleles at a given locus on a particular pair of chromosomes that two relatives inherit from a common ancestor. Because methods based on IBD are more powerful, we shall not consider methods based on marker identity in state, although the two measures of marker similarity approach each other as a marker becomes more polymorphic. Because the estimation of the IBD probability is often separable from the actual regression, the usual regression-based linkage method can be looked upon as a two-stage procedure: (1) the IBD probabilities are calculated; and (2) the regression model is constructed. Obviously, how well we estimate the IBD probabilities at the first stage is directly related to the validity and power of a regression model at the second stage.

The estimation of IBD sharing between two members of a pedigree is based on all of the observed marker genotypic information. Early regression-based methods only used samples of full sibs. If parents are also typed for a codominant marker, or if the parental marker genotypes can be deduced, it is possible to count the number of alleles a sib pair shares IBD. The early methods eliminated those pairs for which it was not possible to actually count the number of alleles shared IBD, and this approach could lead to bias when the marker is not highly informative. Haseman and Elston¹ proposed estimating IBD probabilities at a marker locus by utilizing the marker information available for the sibs and their available parents. The IBD probabilities for other relative pairs can also be accommodated.² Kruglyak and colleagues³ as well as Kruglyak and Lander⁴ proposed using an algorithm (the Lander-Green algorithm) to calculate the IBD allele-sharing probabilities in a multipoint fashion. The amount of computation in their algorithm increases linearly with the number of markers, and exponentially with the size of the family. For large pedigrees, Sobel and Lange⁵ implemented a Markov chain Monte Carlo (MCMC) method and, with the assumption of no interference, Fulker and colleagues⁶ proposed a fast regression method to obtain approximate multipoint estimates of the proportion of alleles shared IBD by full sibs at any location, based on the estimates at the locations for which marker information is available. This regression method was used by Almasy and Blangero⁷ when they implemented a full pedigree variance component likelihood model based on the assumption of trait multivariate normality across pedigree members.

Generally, the informativity of a marker for estimating IBD sharing depends on its degree of polymorphism. For a highly informative marker, the founders of a pedigree have unique alleles and therefore the number of

alleles shared IBD can be unambiguously determined. However, the number of alleles for a marker used in practice is limited, with the result that the number of alleles shared IBD often cannot be specified unambiguously. To describe the degree of polymorphism, two measures are commonly used in practice. One is the marker heterozygosity, which is defined by

$$H = 1 - \sum p_i^2,$$

where p_i is the population frequency of the i allele. From this definition, we can see that heterozygosity is simply the probability that a random individual is heterozygous at a locus in a population with Hardy–Weinberg equilibrium. The other measure is the polymorphism information content (PIC) value.⁸ The PIC value was first derived for a rare dominant disease and it is defined as

$$\text{PIC} = 1 - \sum p_i^2 - 2 \sum_i \sum_{j>i} p_i^2 p_j^2.$$

To measure a marker's informativity for a model-free linkage analysis via pairs of relatives, Guo and Elston⁹ developed a third measure, the linkage information content (LIC). The LIC values measure the probability of being able to determine IBD proportions for each particular type of relative pair.

When the number of alleles shared IBD between members of a relative pair in a pedigree cannot be specified unambiguously, conventionally in a regression-based method we describe the genetic similarity by the estimated proportion of alleles shared IBD in place of the exact proportion of alleles shared. For a sib pair, the proportion of alleles shared IBD — π (which in reality can take on only the values 0, $\frac{1}{2}$, or 1) — is taken to be $\hat{\pi} = \hat{f}_2 + \frac{1}{2}\hat{f}_1$, where \hat{f}_2 and \hat{f}_1 are the estimated probabilities of sharing two and one alleles IBD, respectively, given all of the marker data available. Let f_i be the prior probability that a relative pair shares, by virtue of degree of relationship alone, i alleles IBD. By Bayes' theorem, the estimated probability given the available marker information I_m is simply

$$\hat{f}_i = \frac{f_i P(I_m|i)}{P(I_m)}.$$

When the marker is completely uninformative, we can see that the estimated IBD probabilities are actually the IBD probabilities under the null hypothesis and $\hat{\pi} = \pi_0$, by which we denote the proportion of alleles shared IBD when there is no linkage. This approach has been criticized in the literature because it can lead to loss of power. For example, Schork and

Greenwood¹⁰ pointed out that the estimate of the number of alleles IBD after imputation is biased toward the null and therefore the power can potentially be reduced by less informative markers. To avoid/alleviate this loss of power, some authors have proposed to weight relative pairs according to the informativity of the number of alleles shared IBD.^{10–12} Although weighting approaches to alleviate the loss of power are possible, care should be taken when any such approach is used in practice because the missing mechanism of IBD sharing is not always at random, e.g. when parents are not observed. In this case, the weighting or deleting approach can lead to a worse outcome as it will result in an invalid or conservative statistic. We do not recommend blindly weighting each pair according to the estimated IBD sharing in practice, because the risk is too large compared with the limited possible gain.

Typically, regression-based linkage methods use the estimated IBD proportion, i.e. $\hat{\pi}$, to summarize the marker similarity between relatives in a pair. Kruglyak and Lander⁴ extended the HE method to use the full estimated distribution of 0, 1, or 2 alleles being shared IBD. They argued that the proportion of IBD sharing does not fully utilize the information provided by the whole IBD distribution. Their simulation showed that their modified HE test has better behavior in the presence of uninformative families. The approximation involved in the use of the mean proportion of alleles shared IBD instead of the full distribution has also been tested by Gessler and Xu.¹³ They explicitly make the distinction between the distribution approach and the expectation approach, but they found that there is little difference between them in terms of power. Cordell¹⁴ performed simulations to investigate the test statistics of HE and variance component methods. She found that the expectation approach suffers in both precision and power when the squared difference is the dependent variable in HE regression. However, the simulation parameters in her study were rather extreme because there was no sibling resemblance other than that due to a single QTL, and this QTL explained more than 90% of the trait variance.

As pointed out by Cordell,¹⁴ the best way to minimize the ambiguity of allele transmission when it is difficult to recruit a full sample is to use as densely spaced markers as possible. With the recent availability of a dense map of single nucleotide polymorphism (SNP) markers that can be genotyped automatically, economically, and more accurately, there is great hope that these can be used to improve the current linkage approach to finding genes. However, currently most multipoint linkage programs assume linkage equilibrium among the markers. Whereas this assumption

may be appropriate for sparsely spaced markers with intermarker distances exceeding a few centimorgans, for densely spaced SNP markers linkage disequilibrium (LD) may exist and, if not appropriately allowed for, may lead to bias in the calculation of the probabilities of IBD sharing when pedigrees have missing founder information. It has been shown that when some or all of the parental genotypes are missing, assuming linkage equilibrium among markers when strong LD exists can cause false-positive evidence from ASP data. Because this bias would not affect the correlation between the proportion of IBD sharing and trait similarity measures, the validity of regression-based linkage analysis for a quantitative trait could be robust to this bias, but nevertheless its power may be affected. A simple way to correct this bias at the stage of IBD calculation is to organize the SNPs into nonoverlapping clusters in such a way that one can assume no LD between markers in different clusters and no recombination within each cluster.¹⁵

Another bias in the calculation of the probabilities of IBD sharing, which may be more important in the case of regression-based methods for quantitative trait analysis, is related to population stratification. Although the confounding effect of population stratification on a genetic association study has long been recognized, regression-based linkage methods, like all other model-free linkage analysis methods, are usually thought to be robust to population stratification because the estimation of the probabilities of alleles shared IBD does not depend on the population frequencies of the marker alleles when founder marker genotypes are observed or are deducible. Unfortunately, it is not unusual for founders to be missing in the case of late-onset complex diseases. In this case, the distribution of the estimated IBD sharing may no longer be independent of the marker allele frequencies and, therefore, all linkage studies can be potentially biased by population stratification.¹⁶ For an affected sib-pair design, it has been shown that, when some or all of the founder genotypes are missing, heterogeneity of allele frequencies among the subpopulations can cause excess false-positive discoveries — even when the trait distribution is homogeneous among the subpopulations. After incorporating a control group of discordant sib pairs, or in the case of a quantitative trait, two conditions must be met for population stratification to be a confounder in linkage analysis: the distributions of both the marker and the trait must be heterogeneous among the subpopulations. When this occurs, the bias can result in a test that is either liberal (and hence invalid) or conservative. An obvious way to avoid such deleterious effects from population stratification is to include

as many founders or other family members as possible to reduce the uncertainty about founder marker information.

3. VARIOUS REGRESSION-BASED METHODS WITH DIFFERENT TRAIT SIMILARITY MEASURES

The regression-based linkage analysis method and the maximum likelihood-based variance component methods are two commonly used approaches for the analysis of quantitative traits. The initial regression-based linkage method is the original HE regression¹ for independent full-sib pairs, which has been shown to be robust to selected samples and the distribution of trait values, although it was derived under the assumption of randomly sampled sib pairs and on a squared sib-pair trait difference that is normally distributed. On the other hand, maximum likelihood variance component analysis can be less robust to selected sampling and to nonnormality of the trait values, although it can be more powerful than the original HE regression. Recently, various regression-based methods have been developed with the aim of extending them to any type of family data and improving their power while retaining robustness.

3.1. *The Original Haseman–Elston Regression*

The Haseman–Elston method, as originally proposed by Haseman and Elston,¹ makes use of the squared trait difference between the two sibs in a pair as the measure of trait similarity (in this case, a measure of trait dissimilarity). Let X , the value of a trait, be composed of an overall mean μ , the major genetic effect of a quantitative trait locus (QTL) g , and an independent random effect e . Let $Y_j = (X_{1j} - X_{2j})^2$, where X_{1j} and X_{2j} are the two trait values for the j th full-sib pair. Let π_{tj} denote the proportion of alleles that the j th full-sib pair shares IBD at the trait locus. With the assumption of no dominant effect, it can be shown that

$$E(Y_j|\pi_{tj}) = (\sigma_e^2 + 2\sigma_g^2) - 2\sigma_g^2\pi_{tj},$$

where σ_g^2 is the variance of the QTL and σ_e^2 is the random effect variance.

Let π_{mj} denote the proportion of alleles the j th full-sib pair shares IBD at a marker locus, and \hat{f}_{mij} be the estimated probability that the j th full-sib pair shares i alleles IBD ($i = 0, 1, \text{ or } 2$) at the marker locus, conditional on the marker information available on the sib pair and their relatives. The estimate of the proportion of alleles shared IBD at the marker locus is

$\hat{\pi}_{mj} = \hat{f}_{m2j} + 0.5\hat{f}_{m1j}$. Assuming linkage equilibrium between the marker and trait loci, it follows that

$$E(Y_j|\hat{\pi}_{mj}) = \sum_{\pi_{tj}} \sum_{\pi_{mj}} E(Y|\pi_{tj})P(\pi_{tj}|\pi_{mj})P(\pi_{mj}|\hat{\pi}_{mj}).$$

Then, it can be shown, letting θ be the recombination fraction between the trait and marker loci, that

$$E(Y_j|\hat{\pi}_{mj}) = [\sigma_e^2 + 2(1 - 2\theta + 2\theta^2)\sigma_g^2] - 2(1 - 2\theta)^2\sigma_g^2\hat{\pi}_{mj}.$$

This can be simply written in the form

$$E(Y_j|\hat{\pi}_{mj}) = \alpha + \beta\hat{\pi}_{mj}.$$

Therefore, the hypothesis of no linkage can be tested by a one-sided t -test. Because the regression t -test is derived assuming the residuals of the regression equation are normally distributed, there have been concerns about violation of this statistical assumption. However, it has been shown that the type I error rate is quite robust to deviations from normality for reasonably sized samples. In other words, the regression coefficient, being a weighted average, tends to be normally distributed by reason of the central limit theorem. Wan and colleagues¹⁷ proposed a permutation procedure to evaluate the P -value of the original HE method. The permutation procedure keeps the $\hat{\pi}_{mj}$ in the original order, and randomly permutes the values of Y among sib pairs. For a large number of sib pairs, they showed that the permutation variance of the regression slope and the variance estimated by least squares are equal under the null hypothesis. Simulations showed that the conventional t -test approximates the permutation test quite well. These results indirectly addressed concerns about the assumption violation of the conventional t -test.

Due to its robustness and simplicity, the original Haseman–Elston regression has been widely extended to various situations. Amos and colleagues² proposed a multivariate extension for the original Haseman–Elston method. Tiwari and Elston¹⁸ extended the original HE procedure to two unlinked quantitative trait loci (QTLs) that might interact epistatically. Hanson and colleagues¹⁹ indicated how the original HE test can be readily extended to accommodate parent-of-origin effects, by estimating separate β coefficients according to the parental source of the allele sharing.

3.2. New Regression-based Methods

Compared to the full likelihood-based variance component methods, the original Haseman–Elston regression is less powerful when normality of the

trait is approximately correct. This problem raised the question of how to improve the power of regression-based linkage analysis methods, and invoked a series of papers to make use of a more informative measure of trait similarity. Different trait similarity measures are summarized in Table 1 and are implemented in the software package Statistical Analysis for Genetic Epidemiology (S.A.G.E).²⁰

Wright²¹ initially indicated that using only the squared difference of a pair of trait values as the trait similarity measure may result in loss of some information for linkage, although this had also been noted by Gaines and Elston²² in another context. He pointed out that the full likelihood function for a sib pair can be written in terms of both a sum and a difference. Under the bivariate normal assumption, he demonstrated that a nontrivial amount of power can be gained when the sum of the pair of trait values is also included. Responding to Wright,²¹ Drigalenko²³ first proposed an extension of the HE method that uses both the sib-pair trait sum and difference as dependent variables. Because the squared pair sum of the trait values (taken with opposite sign) results in a regression line that is parallel to that for the squared pair difference, he suggested estimating the slope by simply averaging the estimates from the two regressions of the squared sum and difference, which is the best estimate under the assumption that the residuals have the same variance for both the squared sum and squared difference. He also showed that such a combination is equivalent to performing a single regression using the pair-trait product. Based on the same

Table 1 Definitions of the Dependent Variable for Various Forms of Haseman–Elston Regression

Keyword	Acronym	Dependent Variable	Option in S.A.G.E.
Original	oHE	$-\frac{1}{2}(X_{1j} - X_{2j})^2$	diff
Revisited	rHE	$(X_{1j} - \bar{X})(X_{2j} - \bar{X})$	prod
Weighted ^a	wHE	$\frac{1}{2}[(1-w)(X_{1j} + X_{2j} - 2\bar{X})^2 - w(X_{1j} - X_{2j})^2]$	W2–W4
Sibship sample mean	smHE	$(X_{1j} - \bar{X}_j)(X_{2j} - \bar{X}_j)$	sibship_mean = yes
Shrinkage mean	pmHE	$(X_{1j} - \tilde{\mu}_j)(X_{2j} - \tilde{\mu}_j)$	—

\bar{X} : overall mean; \bar{X}_j : sibship mean; $\tilde{\mu}_j$: shrinkage mean; w : weight.

^aVarious slightly different weighting options are available in S.A.G.E.; W4, which is asymptotically optimal and adjusts for all the nonindependence of full-sib-pair squared sums and differences in larger sibships, is the one used for simulation here.

idea, the overall mean-centered cross-product of sib-pair traits was adopted as the trait similarity measure in the revisited HE method.²⁴

Although the revisited method was presumed to be more powerful than the original HE method, Palmer and colleagues²⁵ showed with an interesting simulation result that the empirical power of the revisited method could be substantially lower than that of the original method when there are strong familial polygenic or environmental correlations. Their simulation result motivated further work by several groups. It is easy to show that the assumption of the same residual variance for the squared sum and difference cannot be attained, and therefore they are not equally informative when there are familial polygenic or environmental correlations. So, the overall mean-corrected cross-product — which weights equally (but with opposite sign) the two slope estimates, i.e. that from the squared difference and that from sum — is not optimal.

Several groups proposed weighting the two slope estimates inversely according to the variances of the squared sum and the squared difference. Let the estimator of the slope from the squared difference be $-\hat{\beta}_1$, and that of the slope from the squared sum be β_2 . Xu and colleagues²⁶ considered a class of estimators for β of the form $w\hat{\beta}_1 + (1-w)\hat{\beta}_2$, where w is a given weight. Let $\hat{\sigma}_{12}$ be the estimated covariance of $-\hat{\beta}_1$ and β_2 , and $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$ be the estimated variances of $-\hat{\beta}_1$ and β_2 . Then, for a large sample size, the estimator β with weight $w = (\hat{\sigma}_{22} - \hat{\sigma}_{11}) / (\hat{\sigma}_{22} + \hat{\sigma}_{11} - \hat{\sigma}_{12})$ has the smallest variance among all linear combinations of $-\hat{\beta}_1$ and β_2 . Visscher and Hopper²⁷ proposed a similar method, which also weights the slope estimates based on the variance estimates separately from the regression models. The method of Forrest²⁸ simultaneously calculates the estimates of the two intercepts and the single slope as well as the two variances, using least squares iteratively. Shete and colleagues²⁹ further proposed a weighting method for larger sibships, allowing for the correlation between pairs within a sibship. Instead of weighting the slope estimates based on the empirical variances from the regression model, the approach of Sham and Purcell³⁰ used variance estimates that they derived as a function of the population trait-pair covariance. They also showed that this method can be used for the selection of maximally informative sib pairs.

Recognizing the loss of power that can result from using the overall sample mean-corrected cross-product as the trait similarity measure, due to the existence of the correlation arising from sibship-specific effects such as common environmental and polygenic effects, investigators proposed using a sibship-specific mean correction to absorb this correlation. Wang and

colleagues,³¹ in particular, proposed the use of a trait product centered by the family-specific sample mean rather than by the whole sample mean. Their simulations showed that centering the trait by the family-specific sample mean results in more power than centering the trait by the whole sample mean when the size of sibship is large and there are family-specific phenotype effects (common environmental and additional QTL effects).

A shrinkage sibship mean-corrected cross-product was also proposed as the trait similarity measure to further improve power.^{32,33} It was shown that the shrinkage sibship mean is actually a weighted average of the overall sample mean and the sibship sample mean, where the weights depend on the variance within and among sibships. When neither the variance within nor among sibships dominates, the shrinkage mean is a combination of both estimators; when the variance among sibships dominates, this trait similarity measure is equivalent to the squared difference used in the original HE regression; and when the variance within sibships dominates, i.e. there is no common environmental or polygenic effect and the QTL effect is not too strong, this trait similarity measure is equivalent to the overall sample mean-corrected cross-product. Simulation results showed that the empirical power of the shrinkage mean-corrected cross-product and the weighted squared sum and difference are similar in most situations.

Another extension of HE regression is the “reversed” method. Sham and colleagues³⁴ noticed that the regression coefficients would be biased when sampling is through the dependent variable. Because it is common to sample subjects according to their trait values, they proposed to avoid biased estimators of the regression coefficient by a regression method that uses the proportion of IBD sharing as the dependent variable and the squared sum and squared difference as independent (predictor) variables. However, in the absence of bivariate normality, there may exist collinearity between the squared sums and differences for pairs within a sibship. To remove this collinearity, they arbitrarily trim the squared differences in such a way that each individual is represented at least once. An alternative is to use both the squares and the cross-products as independent (predictor) variables. This is equivalent to the use of the squared sum and difference, but is more conveniently extended to multivariate phenotypes. When the aim is to test for linkage, Schaid and colleagues³⁵ pointed out that, for the calculation of an LOD score, how the Y value (the transformed trait values in a regression model) is computed from the traits is critical, but whether it is on the left-hand side or right-hand side of the regression equation is not important if there are no other covariates in the regression model. In another

context, Gray-McGuire and colleagues³⁶ noted that, in the presence of other covariates, reversal of the regression equation will, in general, affect the test of significance. More extensive simulations are still needed to examine this “reversed” method.

3.3. Regression-based Linkage Methods for Any Type of Pedigree Structure

The maximum likelihood-based variance component approach, which is based on an assumption about the distribution of trait values across pedigree members, utilizes the whole information from a pedigree in a relatively simple manner. However, the original HE method is based on the assumption of independent full-sib pairs only. In linkage analysis, large sibships and other types of family structures are also often included. It is desirable to extend the original HE to those cases.

For sibships of size larger than two, Blackwelder³⁷ showed that the correlation between pairs of squared sib-pair differences with no sib in common is 0, and that it usually lies between $\frac{1}{4}$ and $\frac{1}{3}$ when there is one sib in common. Assuming multivariate normality of the sibs' trait values, this correlation is $\frac{1}{4}$.³⁸ Single and Finch³⁹ showed that, when sibships of size larger than two are analyzed, a generalized least squares regression can improve the power of the original HE method by allowing for the correlations between the pairs of squared sib-pair trait differences. In the revisited HE method,²⁴ the same approach was adopted. The correlation matrix W of the dependent variable is not an identity matrix, but rather is block diagonal in form with each diagonal block being a matrix W_ρ , where ρ is the correlation between two sib pairs that have one sib in common. Furthermore, it has also been shown that, for a given correlation ρ and sibship size s , the inverse of the correlation matrix can be obtained algebraically, so a generalized least squares estimate can be computed quickly. Shete and colleagues²⁹ adopted this idea to obtain an optimally weighted HE method.

Amos and Elston⁴⁰ extended the original HE method to any type of noninbred relative pair. Let X_j , the value of a trait measured on the j th noninbred individual, be composed of an overall mean μ , a major genetic effect g , and an independently distributed random effect e , with zero mean and variance σ_e^2 . Let π_{ij} denote the proportion of IBD sharing at the trait locus by the j th relative pair, and π_{mj} denote the proportion of IBD sharing at the marker locus. For unilineal relative pairs, we estimate π_{mj} by $\hat{\pi}_{mj} = 0.5\hat{f}_{m1j}$, where \hat{f}_{m1j} is the estimated probability that the j th relative

pair shares one allele IBD. Let $Y_j = (X_{1j} - X_{2j})^2$. Following the same reasoning as for full-sib pairs, it can be shown that for unilineal relative pairs,

$$E(Y_j|I) = \alpha_1 + \beta_1 \hat{\pi}_{mj}.$$

The coefficients of these regressions for each type of unilineal relative pair are given in Table 2. Note that the coefficient β_1 is negative in each case if $\theta < 0.5$ and $\sigma_a^2 > 0$.

Although regression coefficients of various types of relative pairs were derived as early as 1989, an optimal test that combines the information obtained from the various types of relative pairs was not well developed for some time. One of the obvious problems is the dependence among the test statistics derived from the various types of relative pairs. Schaid and colleagues⁴¹ extended the original HE to combine the information from full sibs and half sibs into a single test for linkage. The method estimates the common regression coefficient for full-sib and half-sib pairs, allowing the intercepts and residual variances to differ for full-sib and half-sib pairs. Assuming a multivariate normal distribution, the nondiagonal elements of the variance-covariance matrix, V , was classified into one of six categories, each one being approximated by a function of the variance of the squared difference of full-sib pairs, $V(Y_F)$. A simulation study demonstrated that this approach worked well in many situations, but there are some conditions where the statistic can have a slightly inflated type I error rate.

In linkage analysis, correlated traits of relative pairs in a pedigree form a cluster of correlated observations. In theory, it should be appropriate to use the generalized estimating equation (GEE)⁴² methodology for this situation. Olson and Wijsman⁴³ first considered a HE method to combine various types of relative pairs using GEE. In their paper, the methodology

Table 2 Coefficients of the Regression of Squared Pair Difference on the Proportion of Alleles Shared IBD for Various Relative Pairs

Relative Type	α_1	β_1
Grandchild	$\sigma_e^2 + 2\sigma_g^2 + \theta\sigma_a^2$	$-2\theta(1 - 2\theta)\sigma_a^2$
Half sib	$\sigma_e^2 + 2\sigma_g^2 - 2\theta(1 - \theta)\sigma_a^2$	$-2\theta(1 - 2\theta)\sigma_a^2$
Avuncular	$\sigma_e^2 + 2\sigma_g^2 - \left(\frac{5}{2}\theta - 4\theta^2 + 2\theta^3\right)\sigma_a^2$	$-2\theta(1 - 2\theta)(1 - \theta)\sigma_a^2$
First cousin	$\sigma_e^2 + 2\sigma_g^2 - \left(\frac{4}{3}\theta - \frac{5}{2}\theta^2 + 2\theta^3 - \frac{2}{3}\theta^4\right)\sigma_a^2$	$-2\theta(1 - 2\theta)\left(1 - \frac{4}{3}\theta + 2\theta^2\right)\sigma_a^2$

of GEE was used to provide an estimate of the robust covariance matrix of the set of estimated relative-pair-type-specific regression coefficients. Using this covariance matrix, the asymptotically most powerful test of linkage that optimally combined the information contained in the different types of relative pairs was constructed.

Let W be the set $\{s, g, h, a, c\}$ of subscripts denoting the relative-pair types: (full) sibling, grandparent–grandchild, half sib, avuncular, and first cousin, respectively. Letting Z_{ik} be the squared difference for relative pair i in family k , then

$$E(Z_{ik}) = \mu_{ik} = \alpha_w + \beta_w \hat{\pi}_{ik},$$

where α_w and β_w are parameters that need to be estimated, and $\hat{\pi}_{ik}$ is the estimated proportion of IBD sharing. Let $\lambda = (\alpha_s, \beta_s, \dots, \alpha_c, \beta_c)^T$. An estimate of λ , $\hat{\lambda}$, may be obtained by solving the GEE. Note that $\hat{\lambda}$, conditional on $\hat{\pi}_{ik}$, is consistent and asymptotically normal with a covariance matrix that may be consistently estimated by the robust sandwich estimator. These asymptotic properties hold even if the form of the working covariance matrix is misspecified.

Olson and Wijsman⁴³ discussed possible working covariance matrices, including the independent covariance matrix and a partially specified matrix, and pointed out that some efficiency may be gained by modeling the correlations between relative pairs more precisely. Chen and colleagues⁴⁴ described a more general framework for general pedigrees using GEE. In particular, they showed that the Haseman–Elston methods, the variance component model, and some score tests are all closely related in that the different choices of the working covariance matrix lead to the different methods. Wang and Elston⁴⁵ developed a two-level HE for quantitative trait linkage analysis and general pedigrees under the framework of multiple-level regression. They adopted an iterative generalized least squares (IGLS) algorithm to tractably handle variance–covariance structures varying across families. They showed that the two-level HE can compete favorably with any current version of HE in that it can naturally make use of all the information available in any general pedigree, simultaneously incorporating individual-level and pedigree-level effects and feasibly modeling various complex genetic effects.

4. DISCUSSION

In the literature on QTL mapping, three approaches — regression-based methods, variance component methods, and the newly developed score

statistics — are often used. Although each approach has its own advantages and disadvantages and is usually viewed separately, it is worth noting that they are closely related in large samples because of the similarity of the underlying trait model.

Based on a general trait model, Putter and colleagues⁴⁶ derived a score test for the proportion of the total phenotypic variance due to the quantitative trait locus in a variance component model and showed that, for sib pairs, it is mathematically equivalent to the HE method that optimally combines the squared sum and the squared difference of the centered phenotype values of the sib pairs. Because score tests and likelihood-ratio tests are equivalent for large sample sizes, the variance component likelihood-ratio test is also asymptotically equivalent to this optimal HE test. Their results gave a theoretical explanation of the empirical observations found in simulation studies reporting similar power of the variance component likelihood-ratio test and the optimal HE method.

Based on a trait model in which the trait value is generated by a family-specific effect, a QTL, and a random effect, Tritchler and colleagues⁴⁷ proposed a score test for genetic linkage in nuclear families that applies to any trait having a distribution belonging to the exponential family. They also showed that the score test is closely related to HE methods. For sib pairs, their score test is proportional to the regression estimate of β in the model $\pi_i - 1 = \beta(X_{i1} - \hat{\mu}_i)(X_{i2} - \hat{\mu}_i) + \varepsilon_i$. The HE regression tests are obtained from the above regression by interchanging the response variable and the predictor variable and adding an intercept: $(X_{i1} - \hat{\mu}_i)(X_{i2} - \hat{\mu}_i) = \alpha + \beta(\pi_i - 1) + \varepsilon_i$. Different estimates of $\hat{\mu}_i$ yield different HE test statistics. For example, this test is the original HE regression if $\hat{\mu}_i$ is estimated by the sample mean of a sib pair; while it is the revisited HE regression if $\hat{\mu}_i$ is estimated by the overall mean.

Chen and colleagues⁴⁴ viewed various methods for QTL mapping from the framework of GEE, and different choices of the working matrix lead to the different methods. Although there is a close relationship among the various linkage test statistics for a large sample, the question of which method should be used for a given data set is difficult to answer. The recent computer simulation results from Cuenco and colleagues,⁴⁸ Szatkiewicz and colleagues,⁴⁹ and Chen and colleagues⁵⁰ provide some comparison of the performance of the various new methods in terms of power and the type I error rates.

Here, we have only focused on using the regression framework to detect linkage for a quantitative trait. In the linkage literature, a regression method

is also used in affected sib-pair (ASP) designs to incorporate covariates. Common complex diseases are likely to be genetically heterogeneous, with different genetic and environmental factors contributing to the disease.⁵¹ It is critical to take account of heterogeneity in a linkage analysis. The regression-based methods can be used to deal with heterogeneity in ASP linkage analysis by allowing the IBD sharing probabilities for an ASP to depend on covariate information. For example, Olson⁵² showed that the ASP LOD score can be reparameterized in terms of the natural logarithms of relationship-specific relative recurrence risks. The method incorporates locus heterogeneity by allowing the genetic relative risk to be conditional on indicators of heterogeneity, so that the allele sharing at the marker locus differs for different values of the indicators. The original model of Olson⁵² required two additional parameters for each covariate, and therefore may not be optimal in terms of power. To reduce the number of regression parameters, different approaches have been proposed. Currently, the LOD-PAL software in S.A.G.E. constrains the relative risks in a manner that reduces both the number of parameters in the basic model and the number of additional parameters for each heterogeneity indicator.⁵³

We have discussed many features related to regression-based linkage methods. The power of a regression-based method depends on how we measure the marker similarity and the trait similarity between two relatives in a pair. The regression-based statistics, score test statistics, and variance component statistics are related because they all test for a nonzero QTL variance component. In a general sense, they are all in fact variance component methods, although the usual variance component method is to use a maximum likelihood-ratio test to examine the QTL variance, which is based on the critical assumption of multivariate normality for pedigree data. Apart from being robust to nonnormality, regression-based methods also include the advantage of rapid computation, which makes it feasible to evaluate p -values empirically. The regression framework also offers flexibility to model environmental effects as well as gene-gene and gene-environment interactions.

ACKNOWLEDGMENTS

This work was supported in part by a U.S. Public Health Service Resource Grant from the National Center for Research Resources (RR03655) and a Research Grant from the National Institute of General Medical Sciences (GM28356).

REFERENCES

1. Haseman JK, Elston RC. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**:3–19.
2. Amos CI, Dawson DV, Elston RC. (1990) The probabilistic determination of identity-by-descent sharing. *Am J Hum Genet* **47**:842–853.
3. Kruglyak L, Daly MJ, Lander ES. (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am J Hum Genet* **56**:519–527.
4. Lander E, Kruglyak L. (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* **57**:439–454.
5. Sobel E, Lange K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* **58**:1323–1337.
6. Fulker DW, Cherny SS, Cardon LR. (1995) Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am J Hum Genet* **56**:1224–1233.
7. Almasy L, Blangero J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**:1198–1211.
8. Botstein D, White RL, Skolnick M, Davis RW. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**:314–331.
9. Guo X, Elston RC. (1999) Linkage information content of polymorphic genetic markers. *Hum Hered* **49**:112–118.
10. Schork NJ, Greenwood TA. (2004) Inherent bias toward the null hypothesis in conventional multipoint nonparametric linkage analysis. *Am J Hum Genet* **74**:306–316.
11. Jacobs KB, Gray-McGuire C, Cartier KC, Elston RC. (2003) Genome-wide linkage scan for genes affecting longitudinal trends in systolic blood pressure. *BMC Genet* **4**(Suppl 1):S82.
12. Franke D, Ziegler A. (2005) Weighting affected sib pairs by marker informativity. *Am J Hum Genet* **77**:230–241.
13. Gessler DD, Xu S. (1996) Using the expectation or the distribution of the identity by descent for mapping quantitative trait loci under the random model. *Am J Hum Genet* **59**:1382–1390.
14. Cordell HJ. (2004) Bias toward the null hypothesis in model-free linkage analysis is highly dependent on the test statistic used. *Am J Hum Genet* **74**:1294–1302.
15. Abecasis GR, Wigginton JE. (2005) Handling marker–marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet* **77**:754–767.
16. Wang T, Elston RC. (2005) The bias introduced by population stratification in IBD-based linkage analysis. *Hum Hered* **60**:134–142.
17. Wan Y, Cohen JC, Guerra R. (1997) A permutation test for the robust sib-pair method. *Ann Hum Genet* **61**:7987.
18. Tiwari HK, Elston RC. (1997) Linkage of multi-locus components of variance of polymorphic markers. *Am J Hum Genet* **61**:253–261.

19. Hanson RL, Kobes S, Lindsay RS, Knowler WC. (2001) Assessment of parent-of-origin effects in linkage analysis of quantitative traits. *Am J Hum Genet* **68**:951–962.
20. S.A.G.E. Statistical Analysis for Genetic Epidemiology. Available at <http://darwin.cwru.edu/sage/>
21. Wright FA. (1997) The phenotypic difference discards sib-pair QTL linkage information. *Am J Hum Genet* **60**:740–742.
22. Gaines RE, Elston RC. (1969) On the probability that a twin pair is monozygotic. *Am J Hum Genet* **21**:457–465.
23. Drigalenko E. (1998) How sib-pairs reveal linkage. *Am J Hum Genet* **63**:1242–1245.
24. Elston RC, Buxbaum S, Jacobs KB, Olson JM. (2000) Haseman and Elston revisited. *Genet Epidemiol* **19**:1–17.
25. Palmer LJ, Jacobs KB, Elston RC. (2000) Haseman and Elston revisited: the effects of ascertainment and residual familial correlations on power to detect linkage. *Genet Epidemiol* **19**:456–460.
26. Xu X, Weiss S, Wei LJ. (2000) A unified Haseman–Elston method for testing linkage with quantitative traits. *Am J Hum Genet* **67**:1025–1028.
27. Visscher P, Hopper J. (2001) Power of regression and maximum likelihood methods to map QTL from sib pair and DZ twin data. *Ann Hum Genet* **65**:583–601.
28. Forrest WF. (2001) Weighting improves the “new Haseman–Elston” method. *Hum Hered* **52**:47–54.
29. 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.
30. Sham PC, Purcell S. (2001) Equivalence between Haseman–Elston and variance components linkage analyses for sib pairs. *Am J Hum Genet* **68**:1527–1532.
31. Wang D, Lin S, Cheng R, *et al.* (2001) Transformation of sib-pair values for the Haseman–Elston method. *Am J Hum Genet* **68**:1238–1249.
32. Wright FA. (2003) Information perspectives of the Haseman–Elston method. *Hum Hered* **55**:132–142.
33. Wang T, Elston RC. (2004) A modified revisited Haseman–Elston method to further improve power. *Hum Hered* **57**:109–116.
34. 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.
35. Schaid DJ, Olson JM, Gauderman WJ, Elston RC. (2003) Regression models for linkage: issues of traits, covariates, heterogeneity, and interaction. *Hum Hered* **55**:86–96.
36. Gray-McGuire C, Bochud M, Elston RC. (2007) Genetic association tests based on family data. In: *Proceedings of the Third Seattle Symposium in Biostatistics: Statistical Genetics and Genomics*, Schadt E, Storey JD (eds.), Springer, New York, NY (in press).

37. Blackwelder WC. (1977) Statistical methods for detecting genetic linkage from sibship data. Institute of Statistics Mimeo Series No. 1114, The University of North Carolina at Chapel Hill, Chapel Hill, NC.
38. Wilson AF, Elston RC. (1993) Statistical validity of the Haseman–Elston sib-pair test in small samples. *Genet Epidemiol* **10**:593–598.
39. Single RM, Finch SJ. (1995) Gain in efficiency from using generalized least squares in the Haseman–Elston test. *Genet Epidemiol* **12**:889–894.
40. Amos CI, Elston RC. (1989) Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genet Epidemiol* **6**:349–360.
41. Schaid DJ, Elston RC, Tran L, Wilson AF. (2000) Model-free sib-pair linkage analysis: combining full-sib and half-sib pairs. *Genet Epidemiol* **19**:30–51.
42. Zeger SL, Liang KY. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**:121–130.
43. Olson JM, Wijsman EM. (1993) Linkage between quantitative trait and marker loci: methods using all relative pairs. *Genet Epidemiol* **10**:87–102.
44. Chen WM, Broman KW, Liang KY. (2004) Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman–Elston regression. *Genet Epidemiol* **26**:265–272.
45. Wang T, Elston RC. (2005) Two-level Haseman–Elston regression for general pedigree data analysis. *Genet Epidemiol* **29**:12–22.
46. Putter H, Sandkuijl LA, Houwelingen JC. (2002) Score test for detecting linkage to quantitative traits. *Genet Epidemiol* **22**:345–355.
47. Tritchler D, Liu Y, Fallah S. (2003) A test of linkage for complex discrete and continuous traits in nuclear families. *Biometrics* **59**:382–392.
48. T Cuenco K, Szatkiewicz JP, Feingold E. (2003) Recent advances in human quantitative-trait-locus mapping: comparison of methods for selected sibling pairs. *Am J Hum Genet* **73**:863–873.
49. Szatkiewicz JP, T Cuenco K, Feingold E. (2003) Recent advances in human quantitative-trait-locus mapping: comparison of methods for discordant sibling pairs. *Am J Hum Genet* **73**:874–885.
50. Chen WM, Broman KW, Liang KY. (2005) Power and robustness of linkage tests for quantitative traits in general pedigrees. *Genet Epidemiol* **28**:11–23.
51. Risch N. (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am J Hum Genet* **46**:229–241.
52. Olson JM. (1999) A general conditional-logistic model for affected-relative-pair linkage studies. *Am J Hum Genet* **65**:1760–1769.
53. Goddard KA, Witte JS, Suavez BK, *et al.* (2001) Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. *Am J Hum Genet* **68**:1197–1206.