Composite Statistics for QTL Mapping with Moderately Discordant Sibling Pairs

William F. Forrest and Eleanor Feingold

Department of Human Genetics, University of Pittsburgh

Extreme discordant sibling-pair (EDSP) designs have been shown in theory to be very powerful for mapping quantitative-trait loci (QTLs) in humans. However, their practical applicability has been somewhat limited by the need to phenotype very large populations to find enough pairs that are extremely discordant. In this paper, we demonstrate that there is also substantial power in pairs that are only moderately discordant, and that designs using moderately discordant pairs can yield a more practical balance between phenotyping and genotyping efforts. The power we demonstrate for moderately discordant pairs stems from a new statistical result. Statistical analysis in discordant-pair studies is generally done by testing for reduced identity by descent (IBD) sharing in the pairs. By contrast, the most commonly-used statistical methods for more standard QTL mapping are Haseman-Elston regression and variance-components analysis. Both of these use statistics that are functions of the trait values given IBD information for the pedigree. We show that IBD sharing statistics and "trait value given IBD" statistics contribute complementary rather than redundant information, and thus that statistics of the two types can be combined to form more powerful tests of linkage. We propose a simple composite statistic, and test it with simulation studies. The simulation results show that our composite statistic increases power only minimally for extremely discordant pairs. However, it boosts the power of moderately discordant pairs substantially and makes them a very practical alternative. Our composite statistic is straightforward to calculate with existing software; we give a practical example of its use by applying it to a Genetic Analysis Workshop (GAW) data set.

Introduction

Risch and Zhang (1995) are generally credited with introducing the extreme discordant sibling pair (EDSP) design for quantitative trait–locus (QTL) mapping in humans. The idea of that design is that if phenotyping is much easier than genotyping, then a large population of sibling pairs can be screened to find pairs where one sibling has a very high trait value and the other sibling has a very low trait value. The discordant pairs can then be genotyped to assay identity-by-descent (IBD) sharing at markers. If a marker is linked to the trait, the amount of IBD sharing will be less than would be expected under the null hypothesis of no linkage. The statistical significance of the IBD sharing can be tested with essentially the same "mean sharing" test (e.g., Blackwelder and Elston 1985) that is used with affected sibling pairs for mapping qualitative traits.

The EDSP idea has been further developed in work such as that of Gu et al. (1996), Kruse et al. (1997), Rogus et al. (1997), and Knapp (1998). Most of this literature has focused on designs where one sibling is in the top 10% of the trait distribution and the other is in the bottom 10%, although some work has been done on designs where the second sib is in the bottom 30%, and some work has been done on concordant-pair designs. Despite its potential, the practical applicability of the EDSP method has been somewhat limited by the large number of pairs that must be screened to find pairs that are discordant enough (Risch and Zhang 1996; Nicolaou et al. 1999). One of the few studies using these designs is Xu et al. (1999). They screened over 200,000 adults in Anquing, China in order to ascertain 207 discordant and 357 concordant pairs. Many researchers strapped for time and resources may balk at such figures.

The main alternative to an EDSP study is simply to measure the trait value in a "nonascertained" or "population" sample of families. This requires much less subject recruitment and phenotyping but much more genotyping. The most commonly used analysis methods for this type of study are Haseman-Elston regression (Haseman and Elston 1972) and variance components (Amos 1994). Haseman-Elston regression simply regresses the squared difference in the siblings' trait values on their estimated IBD sharing at a marker. If there is linkage, increased IBD sharing will be associated with

Received July 27, 1999; accepted for publication March 3, 2000; electronically published April 7, 2000.

Address for correspondence: Dr. William F. Forrest, A300 Crabtree Hall, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261. E-mail: forrest@forrest.hgen.pitt.edu

0002-9297/2000/6605-0019$02.00

a smaller squared trait difference and the regression line will have a negative slope. Under the null hypothesis of no linkage, the slope of the regression line is zero. Recently, this method has been updated to use the normalized product of the sibling trait values instead of the squared difference (Wright 1997; Drigalenko 1998; Elston et al. 1999). Variance-components methods, by contrast, rely on estimating variance-components parameters under a Gaussian trait model. They are more frequently applied to larger pedigrees than is Haseman-Elston regression. The covariance matrix of the multivariate trait vector is parameterized through a small number of variance parameters representing various environmental and genetic effects at the locus and the pairwise identity-by-descent relationships of the pedigree members, and maximum-likelihood estimates of the variance parameters are computed at each locus. Testing is done with a likelihood-ratio (LOD score) test, testing the hypothesis that the genetic-variance parameters at the locus are zero. Theoretical references are Amos (1994) and Almasy and Blangero (1998). Examples of applications in the recent literature are Duggirala et al. (1999) and Fisher et al. (1999). Although the variance-components approach is quite different from that of Haseman-Elston regression and its extensions, there is an important unifying feature: both derive their statistical power from examining the distribution of trait values, given the IBD information for the pedigree.

In this paper, we show that IBD sharing statistics such as those used for discordant-sibling-pair designs are statistically independent of "trait value given IBD" statistics such as those used in variance components and Haseman-Elston regression, and thus that the two types of statistics can be combined to give more powerful tests of linkage. This independence result is actually quite general, but we apply it to the special case of discordant sibling pairs. We propose a very simple composite statistic for discordant pairs that just combines the Haseman-Elston regression test statistic with the mean IBD sharing statistic. We examine the power of our composite statistic with simulation studies, and discover that while it increases power only slightly for EDSPs, it gives a great deal of power to moderately discordant pairs. This boost in power means that moderately discordant–pair designs can be a useful middle ground between EDSP studies and population studies, offering a practical balance between the numbers that need to be phenotyped and the numbers that need to be genotyped.

Since our proposed composite statistic is somewhat ad hoc, we discuss two possible refinements to it: weighting the two components unequally, and replacing Haseman-Elston regression with one of its newer extensions or with variance components. We show that different weights are desirable for different ascertainment schemes, and that "original" Haseman-Elston regression is, in fact, probably the best choice for this application. Finally, we point out that our statistic is straightforward to compute using standard software programs, and we demonstrate its application on a data set from the Genetic Analysis Workshop 10 (GAW10). In the discussion we briefly mention other possible applications of our statistical independence result, including statistics for concordant sibling pairs and for experimental crosses.

Theory

In this section, we prove that statistics measuring IBD sharing are statistically independent of those modeling traits conditional on IBD sharing. The section can be skipped by readers interested primarily in applications.

Consider an ascertained sample of n families of a given pedigree structure. For family i, let Φ_i denote the vector of phenotype variables and let Z_i denote the inheritance vector's configuration. The phenotype vector Φ_i may be either a vector of values for family members or a function of those values. For example, in applying Haseman-Elston regression, we ascertain sibling pairs, then define Φ_i to be the squared trait difference rather than the two values themselves. The inheritance vector (Donnelly 1983) is the vector of zeros and ones that tracks which grandparental allele is passed on in each meiosis in the pedigree. Since the number of inheritance vectors is 2^{2A}, where A is the number of pedigree nonfounders, we will typically reduce the inheritance vectors to a collection of $k << 2^{2A}$ equivalence classes known as "inheritance configurations" (see Dudoit and Speed [1999] for a good recent discussion), which can be denoted with arbitrary integer labels $\{1,2,\ldots,k\}$. For $1 \leqslant j \leqslant k$, we use the expression $Z_i = j$ to say that the ith family's inheritance vector at a locus is a member of the jth inheritance configuration. The best known and most widely used example of inheritance configurations is the reduction of the 16 or more inheritance vectors characterizing the potential genetic relationship of two relatives at a locus to the $k = 3$ possible values for the number of genes shared IBD (which has value 0, 1, or 2). IBD sharing between two relatives is a useful example to keep in mind, since both Haseman-Elston and variance-components methods model the pairwise phenotypic dependence for two relatives as a function of the expected number of genes shared IBD at a locus. The pair (Φ_i, Z_i) is the *complete* data for the ith family. Typically, we will observe the phenotypic information (e.g., affected status for a disease, blood pressure, etc.) for each pedigree member, but will get only partial information on the inheritance vector through genotyped markers. We denote the *observed* data by $\mathbf{Y}_\mathrm{i} = (\Phi_i, \mathbf{M}_i = \mathrm{m}(Z_i))$, where $\mathbf{M}_i$ is the marker data available on the *i*th family, which will be some random function m(·) of the inheritance vector and the marker alleles' cardinalities and frequencies.

We assume that some parametric model with parameter vector ψ is specified for the distribution of phenotype Φ_i conditional on inheritance vector Z_i at the locus being tested, with conditional density given by $f_\psi(\phi|z)$. We assume that the distribution of Z_i is multinomial with cell probabilities $\pi = (\pi_1,\dots,\pi_k)$ (e.g., $\pi = (\frac{1}{4},\frac{1}{2},\frac{1}{4})$ for sibling pairs, at an unlinked locus). The joint density of (Φ,Z) is thus the product of the conditional and marginal densities, and can be written

$$f_{\psi,\pi}(\phi,z) = \prod_{j=1}^{k} [f_\psi(\phi|z=j)\pi_j]^{\delta_j} ,$$

where $\delta_j = 1$ when $z = j$ and $\delta_j = 0$ otherwise. For data $\{\Phi_i,Z_i\}_{i=1}^{n}$, we define δ_{ij} such that $\delta_{ij} = 1$ when $Z_i = j$ and $\delta_{ij} = 0$ otherwise. The joint likelihood of a sample of n such families is then

$$\prod_{i=1}^{n} f_{\psi,\pi}(\Phi_i,Z_i) = \prod_{i=1}^{n}\prod_{j=1}^{k} [f_\psi(\Phi_i|Z_i=j)\pi_j]^{\delta_{ij}} .$$

The complete-data log-likelihood, denoted $L^*(\psi,\pi)$, is given by

$$\begin{aligned} L^*(\psi,\pi) &= \sum_{i=1}^{n}\sum_{j=1}^{k} \delta_{ij}\log[f_\psi(\Phi_i|Z_i=j)\pi_j] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \delta_{ij}\log f_\psi(\Phi_i|Z_i=j) + \sum_{j=1}^{k}\log\pi_j\sum_{i=1}^{n}\delta_{ij} \\ &= L^*(\psi) + L^*(\pi) . \end{aligned} \tag{1}$$

Hence, the complete-data log-likelihood separates into two summands containing information about ψ and π separately.

The log-likelihood of the observed data is found by taking the conditional expectation of the complete-data log-likelihood, given the observed data $\{\mathbf{Y}_i\} = \{(\Phi_i,\mathbf{M}_i)\}$, that is, the trait and marker data on each pedigree (Dempster et al. 1977). The difference between the complete-data log-likelihood $L^*(\psi,\pi)$ and the observed-data log-likelihood $L(\psi,\pi)$ is that the 0–1 valued indicator variables $\{\delta_{ij}\}$ are replaced by conditional probabilities of the *i*th family's inheritance vector configuration being of type *j*, denoted $\mathrm{P}(Z_i = j|\mathbf{Y}_i)$. The observed data log-likelihood is therefore

$$\begin{aligned} L(\psi,\pi) &= \sum_{i=1}^{n}\sum_{j=1}^{k} \mathrm{P}(Z_i=j|\mathbf{Y}_i) \times \log[f_\psi(\Phi_i|Z_i=j)\pi_j] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \mathrm{P}(Z_i=j|\mathbf{Y}_i) \times \log f_\psi(\Phi_i|Z_i=j) \\ &\quad + \sum_{j=1}^{k}\log\pi_j\sum_{i=1}^{n}\mathrm{P}(Z_i=j|\mathbf{Y}_i) \\ &= L_1(\psi,\pi) + L_2(\psi,\pi) . \end{aligned} \tag{2}$$

The separation into two pieces—dependent only on ψ and π, respectively—that we demonstrated for the complete-data log-likelihood also holds for the observed-data log-likelihood under the null hypothesis of no linkage (which is the case of practical importance for genome screening). This is because in (2) the first summand $L_1(\psi,\pi)$ depends on π only through the probabilities $\{\mathrm{P}(Z_i = j|\mathbf{Y}_i)\}$, and similarly the second summand $L_2(\psi,\pi)$ depends on ψ only through the probabilities $\{\mathrm{P}(Z_i = j|\mathbf{Y}_i)\}$. Under the null hypothesis of no linkage to the candidate locus, these probabilities depend only on the (fixed) null values of (ψ,π) rather than on their true values at each locus. The assumption of no linkage implies that the phenotypic data is ignored when IBD sharing at a locus is computed and that only the marker data is used. Mathematically, this assumption replaces $\mathrm{P}(Z_i = j|\mathbf{Y}_i)$ by $\mathrm{P}(Z_i = j|\mathbf{M}_i)$ computed under the null hypothesis. Since computing IBD sharing probabilities using the multipoint marker data is the usual method in standard packages such as GENEHUNTER (Kruglyak et al. 1996), SOLAR (Almasy and Blangero 1998), SAGE (Elston et al. 1999), and SIMWALK (Sobel and Lange 1996), in practice, the observed-data log-likelihood implicit in computations separates into summands depending on ψ and π separately, in the form

$$L(\psi,\pi) = L(\psi) + L(\pi) . \tag{3}$$

A more detailed mathematical discussion of these points is contained in Appendix A. However, we stress again that in existing packages, which use only marker data in computing IBD sharing, the observed-data log-likelihood separation holds, giving us $L(\psi,\pi) = L(\psi) + L(\pi)$.

Perhaps the most important implication of the likelihood factorization above is that under regularity conditions the maximum likelihood estimates $\hat{\psi}$ and $\hat{\pi}$ will converge in distribution to independent multivariate normal sampling distributions, so that

$$\mathrm{Cov}(\hat{\psi}_i,\hat{\pi}_j) \approx 0 \;\forall\; i,j . \tag{4}$$

This holds because the Fisher information matrix will be block-diagonal, since

$$I_{i,j}(\psi,\pi) = \mathrm{E}\left[-\frac{\partial^2 L(\psi,\pi)}{\partial\psi_i\partial\pi_j}\right]$$

$$= \mathrm{E}\left\{-\frac{\partial^2[L(\psi)+L(\pi)]}{\partial\psi_i\partial\pi_j}\right\}$$

$$= 0 \ ,$$

implying that its inverse, the normalized asymptotic covariance matrix, will also be block diagonal:

$$I^{-1}(\psi,\pi) = \begin{bmatrix} I^{-1}(\psi) & 0 \\ 0 & I^{-1}(\pi) \end{bmatrix} \ .$$

Independence of $\hat{\psi}$ and $\hat{\pi}$ implies that any functions of them will likewise be independent. Since allele-sharing statistics can be written as functions of $\hat{\pi}$, the implication is that statistics modeling dependence of phenotypes on allele sharing (which typically will be functions of $\hat{\psi}$, in our model's parlance) will be independent of allele-sharing statistics and, hence, can be combined to create more-powerful tests than those based on either phenotype-based methods or allele-sharing alone. Specifically, we show in Appendix B that both Haseman-Elston regression and variance-components methods are convenient approximations of models in the form of the first summand of our observed-data log-likelihood, $L(\psi)$. Thus, our theoretical development implies that Haseman-Elston and its extensions or variance-components statistics can be directly combined with IBD sharing statistics to form more-powerful tests of linkage.

A Simple Statistic for Discordant Sibling Pairs

As an example of the principle that IBD sharing statistics and phenotype-given-IBD statistics can be combined to improve power, we suggest a very simple composite statistic for discordant sibling pairs that essentially just adds the traditional Haseman-Elston and mean IBD sharing statistics. We show by simulation that our composite statistic has higher power than either component alone as long as the sib pairs ascertained are at least moderately discordant.

The Composite Statistic

For discordant sibling pairs, we consider a linear combination of the Haseman-Elston statistic (which stems from the distribution of trait given IBD status) and the mean IBD sharing statistic. If we let $\hat{\beta}_{HE}$ denote the Haseman-Elston slope estimate, and $\hat{\pi}_{n,1} + 2\hat{\pi}_{n,2}$ denote the mean IBD sharing in the n sibling pairs, we define our composite statistic by

$$w_{HE}\frac{\hat{\beta}_{HE}}{\sqrt{\mathrm{Var}(\hat{\beta}_{HE})}} + w_{IBD}\frac{\hat{\pi}_{n,1}+2\hat{\pi}_{n,2}-1}{\sqrt{\mathrm{Var}(\hat{\pi}_{n,1}+2\hat{\pi}_{n,2})}} \qquad (5)$$

so that $w_{HE}^2 + w_{IBD}^2 = 1$; that is, the pair of weights (w_{HE},w_{IBD}) falls on the unit circle. In other words, we standardize each statistic to have a standard normal distribution for n large, then form a linear combination so that the resulting composite test statistic has a standard normal distribution under the null hypothesis (this assumes the components' independence, which relies on theoretical arguments detailed earlier). Use of the Haseman-Elston test alone corresponds to a weighting choice of (1,0); use of the mean IBD test alone corresponds to a choice of (0,1). The optimal weights—those which maximize power—will, in general, depend in complex ways on the interplay of the ascertainment scheme and the unknown genetic model for the disease. However, we will show in simulations and applications that even an ad hoc choice of weights can lead to substantially increased power to detect linkage, because, even if our ad hoc weight choices are not optimal, they will typically be better than the de facto selection of (1,0) or (0,1) made, respectively, when either Haseman-Elston or the mean IBD test is used alone.

Simulation Methods

We simulate sib-pair data from a simple additive QTL model. Let $X_{i,1}$ and $X_{i,2}$ be the trait values for the siblings of pair i, and suppose that

$$X_{i,1} = \mu + g_{i,1} + \epsilon_{i,1}$$

and

$$X_{i,2} = \mu + g_{i,2} + \epsilon_{i,2} \ . \qquad (6)$$

Here μ is the overall trait mean (set to $\mu = 0$), $g_{i,k}$ (for $k = 1,2$) is the genetic effect of the QTL, and $\epsilon_{i,k}$ is environmental noise.

We model the QTL with equifrequent codominant alleles D and d, so that the genetic effect is

$$g_{i,j} = +1 \text{ for a DD individual}$$

$$= 0 \text{ for a Dd individual, and}$$

$$= -1 \text{ for a dd individual.}$$

For purposes of simplicity in the simulation, we further let $\epsilon_{i,k} \sim \mathrm{N}(0,\tau = 1)$ and let the correlation between $\epsilon_{i,1}$ and $\epsilon_{i,2}$, which represents environmental correlation, be 0.20. The genetic variance for this model is $\frac{1}{2}$, and the heritability—defined as the ratio of genetic variance to the sum of the genetic and environmental variances—is $\frac{1}{3}$.

We simulate data from this model for discordant sibling-pair studies, in which we ascertain only siblings for whom one trait value falls into the lower tail of the trait distribution and one into the higher tail. We examine three ascertainment schemes: a population sample, moderate discordance, and extreme discordance. We simulate a population sample of sib pairs for two reasons. This will give us an idea—given our model—of how many sib pairs would need to be ascertained and genotyped to match the power of the discordant designs, which must ascertain (or at least screen) many sib pairs but must contact and genotype relatively few. Also, this simulation acts as a cautionary example of when the composite approach would be very bad indeed. In a population sample, the IBD sharing approach is powerless, so combination of any such statistic with an otherwise useful Haseman-Elston test will result in a degraded linkage-test statistic. We next introduce the concept of moderately discordant sibling pairs (MDSPs). In analogy with EDSPs, a MDSP is one in which the siblings fall into a much larger and more easily attainable set of discordant trait-distribution tails. Here we arbitrarily define these as the lower and upper 35% of the distribution. Finally, we simulate extremely discordant sibling pairs (EDSPs), defined as having one sib trait fall into the top 10% of the population and the other into the bottom 10%.

We ascertain groups under the hypotheses of no linkage ($\theta = \frac{1}{2}$) and of complete linkage ($\theta = 0$) between the QTL and the marker, which is assumed to have eight equally frequent alleles in the population. We simulate only one marker, and so use estimates of $\hat{\pi}$ based only on that marker. We assume the standard normal distribution for the behaviors of our test statistics, so that a specified significance level (we used $\alpha = 0.1\%$, one-sided, throughout our simulations) easily determines a rejection region. Since all our tests for discordant sibling pairs reject for large negative values of the test statistics, our rejection region is $(-\infty, -3.09)$, where -3.09 is the .001 quantile for a standard normal distribution. The data simulated under the null hypothesis of no linkage were used to confirm that the statistics follow the expected standard normal distribution for $\theta = \frac{1}{2}$, and that they have acceptable Type I errors. See table 1.

No linkage between a candidate locus and a QTL implies that $\beta_{HE} = 0$ (i.e., that Φ_i and Z_i are independent) *and* that the siblings should share 0, 1, or 2 genes IBD with probabilities $\{\pi_0,\pi_1,\pi_2\} = (\frac{1}{4},\frac{1}{2},\frac{1}{4})$. We test this joint hypothesis with $\hat{\beta}_{HE}$, $\hat{\pi}_{n,1} + 2\hat{\pi}_{n,2}$ and the composite test (5). We expect that the composite will be more powerful than either Haseman-Elston or allele sharing alone whenever the latter two both have some reasonable power to detect linkage. Our purpose in examining this statistic is to demonstrate that very simple combinations of existing statistics can yield improvements over standard methods in appropriately ascertained samples.

Table 1

Empirical Type I Error of Three Linkage Tests

Sample Size	Haseman-Elston	Mean IBD	Composite
Population sample:			
200	.0012	.0003	.0006
500	.0005	.0005	.0007
800	.0010	.0003	.0003
1100	.0009	.0002	.0007
1400	.0012	.0006	.0006
MDSPs:			
200	.0013	.0005	.0006
250	.0010	.0006	.0006
300	.0013	.0002	.0008
350	.0006	.0002	.0011
400	.0007	.0005	.0008
EDSPs:			
25	.0022	.0002	.0003
50	.0015	.0005	.0003
75	.0010	.0007	.0004
100	.0010	.0007	.0007
125	.0014	.0003	.0004

Simulation Results

We display in figure 1.1.a the approximate power (on the basis of 10,000 simulations and under the assumption of a one-sided significance level of $\alpha = 0.1\%$) for sample sizes ranging from 200 to 1400 sibling pairs, under the assumption of random ascertainment from the population. As we expect, the mean IBD test of linkage is useless here: Since the sib pairs are randomly picked from the population without regard to trait values, their IBD status at the QTL follows the $(\frac{1}{4},\frac{1}{2},\frac{1}{4})$ proportions expected under no linkage. The most powerful of our three tests is Haseman-Elston. The composite test performs noticeably worse than Haseman-Elston. This is expected, since it combines the useful information which constitutes the Haseman-Elston test with the random noise of the mean IBD statistic and, hence, loses power. The ideal weights under this scenario are $(w_{HE},w_{IBD}) = (1,0)$.

Figures 1.1.b and 1.1.c show the joint distribution of the normalized Haseman-Elston slope estimates and normalized mean IBD statistics under the hypotheses of no linkage and complete linkage, respectively, for a sample size of 200 sib pairs. We see that the joint distribution from data simulated with no linkage (figure 1.1.b) appears to have uncorrelated components with means about the origin. With linkage, the components still appear uncorrelated, but the overall mean of the Haseman-Elston component shifts away from zero, while the mean IBD statistic retains an overall mean near zero (figure 1.1.c).

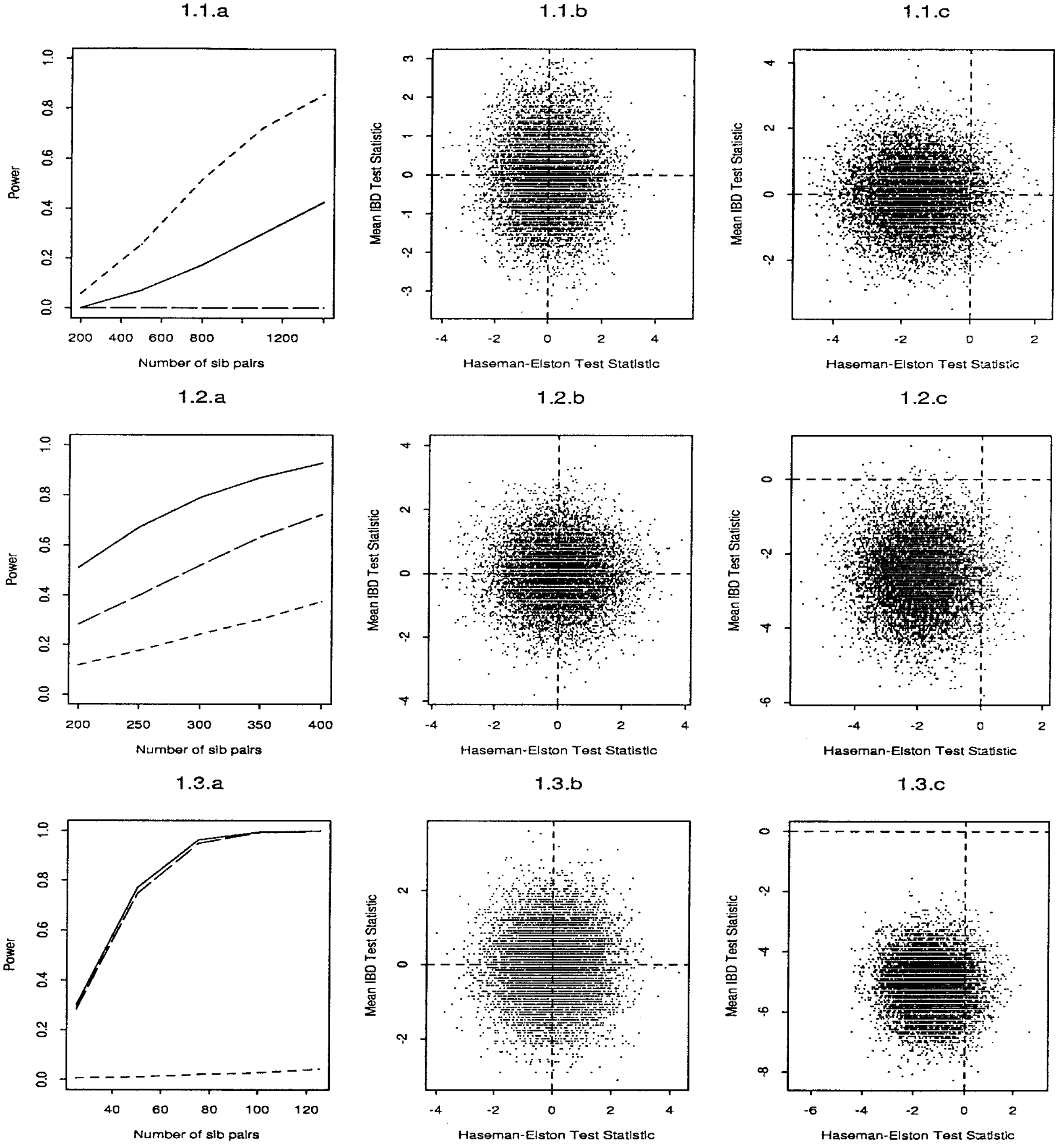

Figure 1 Power curves and scatter plots of the component test statistics over 10,000 replications for the simulation experiment. 1.a–1.1.c refer to the linkage analysis with a population sample of sib pairs. *1.a,* power curves when $\theta = 0$ for the Haseman-Elston test (- - - - - - -), the mean IBD test (— — — —) and the composite test (————) over the five sample sizes considered. *1.b* and *1.c*, paired values for the Haseman-Elston test statistic and the mean IBD test for the cases $\theta = \frac{1}{2}$ and $\theta = 0$, respectively, for a population sample of 200 sib pairs. 2.a–2.c show the corresponding plots for MDSP samples. In 2.b and 2.c, 200 MDSPs are analyzed in each of the 10,000 replications for $\theta = \frac{1}{2}$ and $\theta = 0$, respectively. 3.a and 3.c show the corresponding plots for EDSP samples. In 3.b and 3.c, 100 EDSPs are analyzed in each of the 10,000 replications for $\theta = \frac{1}{2}$ and $\theta = 0$, respectively.

Figure 1.2.a shows power (for $\alpha = 0.1\%$, one-sided) computed for a range of 200 to 400 moderately discordant sibling pairs, so that one has a trait value below the 35% quantile, while another has a trait value above the 65% quantile. On average, we screen 6.2 sibling pairs in our simulation to find each such MDSP. Here Haseman-Elston is less powerful than the mean IBD statistic, and both are less powerful than the composite statistic with ad-hoc equal weights $(w_{HE},w_{IBD}) = (\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$.

Figures 1.2.b and 1.2.c show the joint distribution of the Haseman-Elston slope estimates and mean IBD statistics in the case of moderate ascertainment for $\theta = \frac{1}{2}$ and $\theta = 0$, respectively. With selective ascertainment, the mean IBD statistics derived from data simulated assuming linkage are not centered about zero anymore, as they were in figure 1.1.c. Note also that we can see here empirically that the Haseman-Elston slope estimates (components of $\hat{\psi}$ in our theoretical derivation) are uncorrelated with the mean IBD estimates (functions of $\hat{\pi}$), as we expect on the basis of (4), and that the lack of correlation holds for both no linkage and complete linkage in the simulated data.

Finally, we consider the approach of ascertaining extremely discordant sibling pairs in QTL analysis. Here we require that one sibling fall below the 10% quantile and the other above the 90% quantile in order to be included in the linkage study. This approach is well-known to be very powerful, but can be difficult to implement because so many families must be screened in order to find the discordant siblings. In our simulation, over 158 families are screened on average to find each discordant sibling pair. The figure 158 is specific to our study, but in genetic models for which the discordant sibs are easier to find, they are often less powerful, so the problem is not easily circumvented. Risch and Zhang (1996) discuss this issue in more detail.

Figure 1.3.a shows approximate power curves (again for a one-sided significance level of $\alpha = 0.1\%$) for our three tests, when the numbers of subjects range from 25 to 125. Since we expect that the IBD sharing statistic will be far more powerful here than the Haseman-Elston test, equal weights are inappropriate. For EDSPs, we propose as a conservative ad-hoc choice to use

$$(w_{HE},w_{IBD}) = \left[\cos\left(\frac{5\pi}{12}\right),\sin\left(\frac{5\pi}{12}\right)\right] \approx (.259,.966) \ .$$

We discuss this choice and other issues of weighting in a subsequent section. The mean IBD statistic is far more powerful than Haseman-Elston here, so that the improvement of the composite statistic over the mean IBD test with this model is very modest. Nonetheless, the Haseman-Elston component represents useful data which would otherwise be discarded, so that there seems to be little reason *not* to take the slender advantage offered by a composite statistic, given the amount of time and effort going into a typical linkage study.

The joint distributions for the Haseman-Elston slope and mean IBD estimates under the null and alternative hypotheses are displayed in figures 1.3.b and 1.3.c, respectively, for a sample size of 100 sibling pairs. We see once again that under the hypothesis of no linkage, the normalized test statistics in repeated studies appear to be uncorrelated and scattered about the origin. Under complete linkage, and with highly selective discordant ascertainment, the lack of correlation remains, but the mean IBD statistics are centered far below zero. The Haseman-Elston slope estimates are also centered below zero, though their center appears comparable to what we saw with less selective ascertainment.

Application to GAW 10 Data

To demonstrate a practical application of our method, we apply it to some data from the tenth Genetic Analysis Workshop (GAW). This is very high-quality simulated data, so that the genetic etiology of the observed quantitative traits will be known, but the situation has the complexity of a real study. We focus on nuclear family data from problem 2A of GAW 10. The scope of the simulation is too extensive to recount here, so we refer the reader to MacCleur et al. (1997) for details, and instead use a small piece of this complex data set to illustrate our method. For demonstration purposes, we arbitrarily pick the fourth quantitative trait, which is a complex trait influenced by several genes including MG4, located on chromosome 8. We attempt to detect and locate the gene on chromosome 8 which contributes to the trait.

A great deal of previous work has been done on the GAW 10 data set using both trait-given-IBD methods (see, e.g., Amos et al. 1997; de Andrade et al. 1997; Pugh et al. 1997) and allele sharing statistics (see, e.g., Rogus et al. [1997]; Gu et al. [1997]; Kruse et al. [1997]; and Rochberg et al. [1997]). We combine these two lines of inquiry by ascertaining families with moderately discordant sibling pairs and analyzing them with our composite statistic. The composite statistic based on our theoretical independence result is what makes ascertaining MDSPs a sensible strategy for QTL analysis. Moderately discordant sibling pairs are relatively easy to find (compared to EDSPs), require the genotyping of only a subset of those phenotypically screened, and yield evidence of linkage in two independent ways: through trait-given-IBD statistics (e.g., Haseman-Elston regression) and through IBD-sharing statistics (e.g., the mean-IBD test). This approach provides solid evidence for linkage with much smaller populations than are usually required for discordant sib-pair analyses. The MDSPs ascertained are, of course, less informative than extremely discordant ones. They are, however, much easier to find, and we demonstrate that the LOD score at the QTL is increased by adding in information from a trait-given-IBD statistic.

To demonstrate the moderately discordant ascertainment scheme, we take as a population the first three independent replicates of 239 nuclear families, for a total population of 717 nuclear families. In each family, we take the most discordant pair of sibs such that one falls into the lower 35% of the quantitative-trait distribution and one into the upper 35%. Such an ascertainment scheme yields 204 moderately discordant pairs from the first 717 families screened. We analyze these pairs using MAPMAKER/SIBS (Kruglyak and Lander 1995). We first scan chromosome 8 for linkage using Haseman-Elston regression and store the resulting normalized slope estimates in a file. Next, we compute multipoint estimates of IBD sharing for each sib pair using output from the "dump ibd" command line option. Finally, we compute the normalized test statistics from both Haseman-Elston and mean IBD sharing, then use equal weights $(w_{HE},w_{IBD}) = (\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$ to compute the composite statistic using both tests, and convert all three to lod scores for plotting.

The results are shown in figure 2. We see that the LOD score of the composite test near the MG4 locus on chromosome 8 is highly significant (exceeding 3.8), although neither Haseman-Elston nor the mean IBD test has a LOD score in excess of 2.71, evidence of linkage which in a real genome scan probably would be regarded as inconclusive at best and perhaps would be ignored altogether. Furthermore, the peak of the composite test curve is closer to the MG4 locus than is that of its most powerful component, the Haseman-Elston curve. This occurs because although the Haseman-Elston curve has two peaks of near-equal height to the right of the true locus, the independent mean IBD statistic has its peak—albeit a low peak—precisely over the MG4 locus. In this instance, the mean IBD test is much less powerful than Haseman-Elston, but it serves both to greatly increase the signal at the gene locus and to dampen the signal (relative to the true peak) at false peaks. Finally, we note that assigning equal weights is probably not optimal in this case. Nonetheless, the ad-hoc assignment of equal weights markedly improves performance over the component statistics, presumably because it is closer to the (unknowable) correct set of weights than is the choice of $(w_{HE},w_{IBD}) = (1,0)$ or $(w_{HE},w_{IBD}) = (0,1)$, inherent in the application of Haseman-Elston or the mean IBD test, respectively. This demonstrates that suboptimal and arbitrary weights still

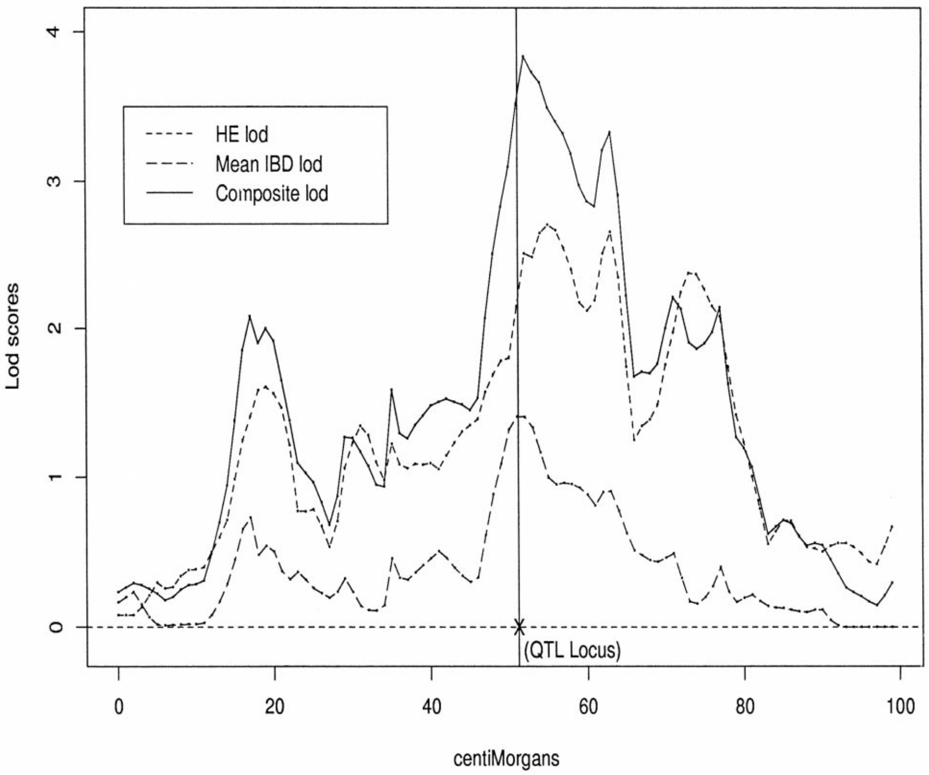

Figure 2 LOD scores for the three methods, computed at 1-cM spacing along chromosome 8 for 204 moderately discordant sibling pairs taken from the first 717 (i.e., three replications) GAW10 problem 2A nuclear families. Ascertainment is such that one sib falls into the top 35% of the trait distribution and the other into the bottom 35%.

Table 2

Power Comparison of Old Haseman-Elston, New Haseman-Elston, Weighted Haseman-Elston, and Variance Components for the Three Ascertainment Plans (Population Sample, Moderate Discordance, and Extreme Discordance) Used in the Simulations

Ascertainment Scheme	Old Haseman-Elston	New Haseman-Elston	Weighted Haseman-Elston	Variance Components
Population	.2220	.2097	.2971	.2988
Moderate discordance	.3582	.3588	.0009	.0023
Extreme discordance	.3230	.3310	.0018	.0022

NOTE.—Sample size is fixed at 200, $\theta = 0$, and $\alpha = 1\%$ one-sided.

lead to substantially better performance in practical situations.

In summary, both Haseman-Elston and the mean IBD test provide some evidence of linkage, although it is not "conclusive" (i.e., it does not have a LOD score >3) for either method. Our independence result and composite statistic combine these two tests to make a stronger case for linkage, while using only 204 sibling pairs ascertained from a population of 717 nuclear families.

Improvements to the Composite Statistic

In this section we consider two questions. First, is Haseman-Elston regression a good choice for the trait-given-IBD component of our test statistic? Second, what are optimal—or at least good—choices of weights for different ascertainment schemes and genetic models?

Is Haseman-Elston the best choice for this composite statistic?

Since the advent of the Haseman-Elston method (Haseman and Elston 1972), which regresses the sibs' squared trait difference on their mean IBD status at a locus, a number of more sophisticated techniques have been introduced for QTL detection in sib pairs. In this section, we briefly review some of these methods, then present simulation results to justify our choice of Haseman-Elston regression. All the methods we discuss in this section model sib traits conditional on sibs' IBD status, so that the resultant test statistics will all be uncorrelated with IBD sharing statistics such as the mean IBD test.

Variance-components analysis (Amos 1994; Almasy and Blangero 1998) has been applied to sib-pair data (Fisher et al. 1999) and has been found to be quite powerful (Fulker and Cherny 1996), though not as computationally simple as Haseman-Elston regression. Wright (1997) opened the door to improvements in the Haseman-Elston method by pointing out that the sib traits' difference and sum conditional on IBD status are statistically independent and, hence, can provide complementary information for linkage analysis. Drigalenko (1998) and Elston et al. (in press) attempt to apply Wright's findings by regressing the sibs' centered trait product on IBD status, a method we will refer to as "new Haseman-Elston." Recent work by W. F. Forrest (unpublished data) demonstrates that new Haseman-Elston regression can lose substantial power because it fails to weight the contributions of the squared trait difference and mean-corrected sum by their disparate variances. The mathematical forms of new Haseman-Elston and Forrest's extension, which we call "weighted Haseman-Elston," are sketched in Appendix C and will be discussed in more detail in future work.

We compared the old Haseman-Elston, new Haseman-Elston, weighted Haseman-Elston, and variance components under our simulation model and three ascertainment schemes, using 10,000 replications each for a sample size of 200 sibling pairs. The old Haseman-Elston is the simplest, best known, and most widely implemented of these methods. Based on simulations, it is also about as powerful as any of these methods in applications to discordant sib-pair data. Table 2 shows the relative power of these methods to achieve significance at the $\alpha = 1\%$ level, one-sided, for the three ascertainment schemes we used in our simulation.

In a population sample, we see that variance-components analysis and weighted Haseman-Elston both provide high power compared to old and new Haseman-Elston. Old Haseman-Elston appears to be at least as powerful as the new Haseman-Elston method. This fact is due to the residual (i.e., environmental) correlation of 0.20 included in our QTL model. The relative power of old and new Haseman-Elston can be shown to depend heavily on this residual correlation.

In applications to both MDSPs and EDSPs, however, we see that old Haseman-Elston does at least as well as any of the other methods. The very low power of weighted Haseman-Elston and variance components stems from their heavy dependence on the distributional assumption of approximate bivariate normality in the sib-pair trait distribution. By ascertaining discordant sibs exclusively, we dramatically violate this assumption. The sibling pairs observed do not come from anything remotely resembling the approximate bivariate normal trait distribution implicit in the QTL model (6) and exemplified in figure 3a. By selecting, for instance, MDSPs, we induce a very peculiar sampling distribution in the observed traits, embodied in figure 3b. The ascertained sib pairs have traits which are strongly negatively correlated and do not follow a bivariate normal distribution. The effect on weighted Haseman-Elston and sib-pair variance components is catastrophic, since both are derived from a bivariate normal trait-given-IBD model with correlations increasing with increasing IBD status. Both methods fail badly as a result. On the basis of this simulation, we recommend against analyzing discordant sibling-pair data with either weighted Haseman-Elston or variance components. In discordant sib pairs, the fact that old Haseman-Elston regression does not use all the data in the bivariate distribution (Wright 1997) works to its advantage, since the unused information seems to be, in part, what is lost during the ascertainment process.

Given these findings and the widespread implementation of and familiarity with old Haseman-Elston, the method seems a reasonable choice for the trait-given-IBD component of our composite statistic. We reiterate, however, that any other trait-given-IBD statistic (such as new Haseman-Elston) can be used and the two components will be independent.

Refining the Weights

One thing that is "simple" about our simple composite statistic is that we combined the two components with equal weights. There is no reason to expect this to be optimal in general, and we now present a small investigation into picking sensible weights for a variety of models and discordant ascertainment schemes.

The weights (w_{HE},w_{IBD}) are constrained to fall on the unit circle (so that $w_{HE}^2 + w_{IBD}^2 = 1$), because this leads to a convenient distribution for the composite statistic. Since the two components are assumed to be normalized so that, under the null hypothesis, they both have an expected value of zero and unit variance, the unit circle convention ensures that the composite statistic likewise follows a standard normal distribution at unlinked loci. Given the unit circle constraint, the weights can be summarized into a single number, the angle measured from the negative x axis to the weights in the third Cartesian quadrant. For this angle θ, the optimal weights are $(w_{HE} = \cos(\theta),w_{IBD} = \sin(\theta))$.

For the two ascertainment schemes shown in our simulations above, the empirically determined optimal weights are (0.615, 0.788) for the moderate ascertainment and (0.256, 0.969) for the extreme ascertainment, corresponding to respective angles of 52° and 75°. We perform further studies to determine the optimal weights over a broader range of genetic parameters and ascertainment schemes. We experiment with QTL frequencies of 0.20 and 0.50, residual correlations of 0.20 and 0.40, and heritabilities of 0.15 and 0.333. We then examine these $2^3 = 8$ models under four ascertainment schemes. First, we ascertain a sibling pair if one sib falls above the median and one below. We call this the (.50, .50) scheme. Next we ascertain a sib pair for which one falls into the top 35% and one into the bottom 35% of the distribution, which we call the (.35, .65) scheme. Similarly, we consider a (.25, .75) scheme and an extreme (.10, .90) scheme. In each case, we screen two-child nuclear families until we have 200 sib pairs, then compute the Haseman-Elston test and the mean IBD test. We re-

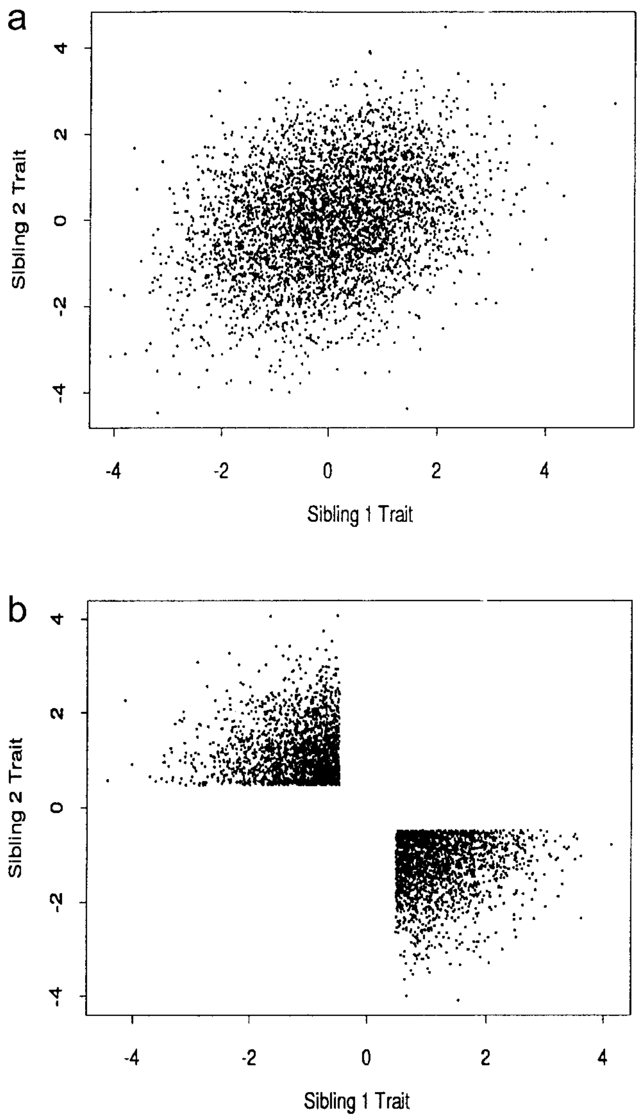

Figure 3 *a,* bivariate scatter plot of 5,000 trait pairs for simulated sib pairs ascertained at random from the population. The points are well described by a bivariate normal distribution. *b,* bivariate scatter plot of 5,000 trait pairs for simulated sib pairs ascertained subject to the restriction that one trait fall into the lower 35% of the trait distribution and the other into the upper 35% of the distribution. The points are restricted to two diametrically opposed regions because of the ascertainment scheme and, thus, are not well described by a bivariate normal distribution.

peat this process 10,000 times and use the results to estimate the optimal weights for that scenario.

We present the results in table 3. Shown are point estimates and 95% confidence intervals for the angle (measured from the negative *x* axis) denoting optimal weights. For reference, our simulations' choices of $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ for moderate discordance and $(\cos(\frac{5\pi}{12}), \sin(\frac{5\pi}{12}))$for extreme discordance correspond to angles of 45° and 75°, respectively. These weights could also be found by bivariate numerical integration (similar to the work in Risch and Zhang 1996), but we found simulation—which amounts to Monte Carlo integration—to be sufficiently accurate and much more convenient.

The most heartening result in table 3 is that the weights do not vary much within an ascertainment scheme. This is very important, because the way that data are ascertained is most often known, but the genetic model is not. Note that this does *not* imply that the component tests perform in similar ways across all genetic models. In fact there are large changes in the power to detect the QTL as we change the allele frequency, residual correlation, and heritability. However, these changes tend to happen in roughly the same way for both Haseman-Elston regression and the mean IBD test, so that the weights remain stable within an ascertainment scheme. For example, if we increase heritability from 0.15 to 0.33, keeping everything else the same, the power of Haseman-Elston increases, but so does the power of the mean IBD test. Since our weights measure the relative strength of each component, they apparently undergo only small changes across the models tested. This means that a guess at ad hoc weights can be based on the ascertainment scheme and expected to be nearly correct for a variety of genetic models.

We note, for completeness, that the ideal weights depend marginally on the quality of the IBD information—that is, on the informativeness and density of the markers. This is because the mean IBD sharing statistic is normalized by a standard error that assumes perfect IBD information (see Kruglyak et al. 1996). When the IBD information is not perfect, the normalized statistic has a variance slightly smaller than 1, making it slightly less powerful than it would be if we could normalize it by the correct variance. (This effect also explains why, in table 1, the mean IBD test's type I errors are slightly, though consistently, smaller than $\alpha = .0010$.) The loss of power is reflected in the weights, which measure how powerful the IBD sharing statistic is relative to the Haseman-Elston statistic. Our simulations use a single marker with eight equifrequent alleles, which gives an overall level of information similar to what is seen with a 10-cM map of typical microsatellite markers. If we had had perfect IBD information, the angles would have come out slightly larger than those shown in table 3. If we had had very poor IBD information (e.g., a single marker with only a few alleles), the angles would have been slightly smaller. We reran a few simulations with different levels of IBD information and saw changes in the angle of roughly 1°–3°. We do not expect this to make much practical difference.

Discussion

We have described a theoretical framework unifying two frequently applied approaches to QTL linkage analysis. The foundation of our approach is the distinction between the inheritance vector distribution and the conditional distribution of phenotypic trait given the inheritance pattern. We observe that the two distributions can carry complementary information, because the null

Table 3

Simulation-Based Optimal Weights for a Variety of Discordant Ascertainment Schemes and Genetic Parameters

QTL Frequency	Residual Correlation	Heritability	Ascertainment Scheme Tail Quantiles (lower, upper)			
			(.50, .50)	(.35, .65)	(.25, .75)	(.10, .90)
.2	.2	.15	33.7 ± 1.3	48.6 ± 0.9	58.1 ± 0.7	71.1 ± 0.4
.2	.2	.33	34.6 ± 0.5	50.1 ± 0.4	61.0 ± 0.3	73.9 ± 0.2
.2	.4	.15	31.1 ± 0.9	49.2 ± 0.6	59.4 ± 0.5	73.6 ± 0.2
.2	.4	.33	33.1 ± 0.4	51.6 ± 0.3	63.0 ± 0.2	75.6 ± 0.1
.5	.2	.15	35.2 ± 1.2	50.5 ± 0.9	59.9 ± 0.6	72.3 ± 0.4
.5	.2	.33	35.8 ± 0.5	52.1 ± 0.4	62.5 ± 0.3	75.7 ± 0.2
.5	.4	.15	32.6 ± 0.9	49.8 ± 0.6	61.1 ± 0.5	74.4 ± 0.3
.5	.4	.33	33.5 ± 0.4	52.9 ± 0.3	63.9 ± 0.2	77.5 ± 0.1

NOTE.—We consider QTL frequencies of .2 and .5, residual correlation of $\rho = .2$ and $\rho = .4$, heritability values of $h = 0.15$ and $h = \frac{1}{3}$, and discordant ascertainment schemes in which the lower and upper tails range from very moderate (.50, .50) to extreme (.10, .90). Presented are 95% confidence intervals for angles measured in degrees from the negative *x* axis into the third Cartesian quadrant. For an angle θ, the optimal weights are given by $w_{HE} = \cos(\theta)$ and $w_{IBD} = \sin(\theta)$.

hypothesis of no linkage implies two separate conditions:

1. The inheritance vectors are uniformly distributed. This allows calculation of the null distribution of any function of the inheritance vectors. For example, full siblings should share 0, 1, or 2 genes IBD in the familiar proportions of ($\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$) at a locus unlinked to the gene controlling the ascertainment-determining phenotype.

2. Phenotypic traits are independent of the inheritance vectors at an unlinked locus, so that the distribution of trait conditional on inheritance vector is simply the marginal distribution of the trait, or $f_\psi(\phi|z) = f_\psi(\phi)$.

Traditional allele-sharing methods test for deviations from the first condition, while Haseman-Elston regression, its extensions, and variance-components methods test for deviations from the second. We posit in (1) and (2) a likelihood function which combines evidence from both kinds of deviation simultaneously. We derive in Appendix B an approximation that includes common methods such as Haseman-Elston regression and variance-components analysis as special cases of the first half of our likelihood function. The observed data log-likelihood itself separates in practice as $L(\psi,\pi) = L(\psi) + L(\pi)$, so that the maximum-likelihood estimates of model parameters ψ and π will be approximately independent. We applied this independence result to create a composite statistic that can increase the power to detect linkage using discordant sibling pairs.

Our simulation results show that the composite statistic increases power only very slightly for EDSPs chosen according to the usual definition of one sibling in the top 10% of the trait distribution and one in the bottom 10%. However, when applied to MDSPs (one sibling each in the top 35% and bottom 35%), the composite statistic increases power enormously. This increase in power means that MDSPs may be a very sensible choice for many linkage studies. For the trait model used in our primary simulation, achievement of 80% power with EDSPs required that ~8,700 pairs be screened to ascertain 55 subjects. For MDSPs, we needed to screen only 1,850 and genotype 300. With a population sample, we needed to ascertain and genotype about 1,300 pairs with Haseman-Elston regression. Although this number drops to about 950 if variance components or weighted Haseman-Elston is applied (as is likely with a population sample), it still represents a more than three-fold increase in the amount of genotyping required over that in MDSP analysis. Our choice of the (.35,.65) criterion as our definition of an MDSP was somewhat arbitrary. In any particular study, a more or less extreme criterion could yield a more convenient ratio of phenotyping to genotyping effort. We did not make any attempt to describe an "optimal" criterion, since that would depend on the relative costs of screening/phenotyping and genotyping.

Overall, we recommend the use of our composite statistic whenever discordant sibling pairs are being used for QTL mapping. For EDSPs, it is likely that the benefit will be small, but there is no reason not to go ahead and gain that small advantage. The two parts of the statistic can easily be computed with standard packages, and then weighted and added. For our (.35, .65) criterion, equal weights seem very close to optimal under a variety of trait models. Table 3 suggests weights for other ascertainment schemes. These weights are specific to our use of the original Haseman-Elston statistic and so would not necessarily be appropriate if a different "trait given IBD" statistic were used. However, we stress that, despite the availability of newer competitors, we found Haseman-Elston regression to be the best choice for this application.

Although all of our discussion has focused on discordant sibling pairs, other applications of our independence result are possible. The most immediately obvious extensions are to concordant sibling pairs and to mixed concordant and discordant designs. Studies of concordant siblings can arise either because the siblings are ascertained explicitly on the basis of their quantitative phenotypes or—more frequently in the study of sib pairs—because they are "affected" by some condition so that both have unusually high or low values for some associated quantitative trait. A recent example of an affected-sib-pair study with concordant quantitative-trait data is the investigation of Type 2 diabetes in Finnish sib pairs by Ghosh et al. (1999). In a concordant-sibling-pair design, the approach we outlined carries over with few changes. The "trait given IBD" component should be one for which the phenotype function varies noticeably as a function of IBD status. Since trait differences for concordant sib pairs are conditioned to have little variability (this is, after all, the definition of concordance), standard Haseman-Elston regression is not a good choice. Taking a cue from Wright (1997), we suggest regressing the mean-corrected squared trait sum on expected IBD status, and looking for a significantly positive slope to indicate linkage. A sensible composite test is then a linear combination of the mean IBD test and the normalized regression slope of squared trait sum given IBD status. It may also be reasonable to apply variance-components methods as the trait-given-IBD component, though we have not explored potential composite statistics for this choice in any detail.

If concordant and discordant pairs are used together, as in the EDAC design of Gu et al. (1996), then Haseman-Elston and trait-sum regression would be carried out separately within the discordant and concordant sibs, respectively. There would be four weights to assign in such a statistic, but we have not investigated what those should be. We note that just as the mean IBD test is generally less powerful for concordant sib pairs than for discordant ones (Risch and Zhang 1995), squared trait sum regression is less powerful than Haseman-Elston regression (unpublished data). The upshot is that although a composite test based on concordant sib pairs will typically be more powerful than one based on either concordant component alone, for a given sample size, it may still be less powerful than one based on discordant sibs. On the other hand, concordant sibs are easier to find than are discordant sibs for many quantitative traits, so ascertaining a larger number of them may not be a problem.

A less obvious but potentially useful application of our independence result is to studies that use larger pedigrees ascertained because numerous family members have extreme trait values. Standard variance-components methods could provide a linkage test based on the distribution of phenotype given IBD status. For an IBD sharing statistic, promising options are statistics intended for linkage studies of qualitative (binary) traits, such as S_{pairs} (see McPeek 1999 for a recent discussion of this and other such statistics). The two statistics could then be combined into a composite, as in our discordant-pair work. The composite might in theory increase power, although the practical details need to be investigated carefully.

Our work here may facilitate similar composite QTL statistics in some experimental crosses. For a simple example of how this might be done, consider a standard backcross between a pure line (A) and a hybrid line (H) with different trait distributions, resulting in offspring which are either A or H at each marker. Suppose further that the pure line (A) is susceptible to a rare condition of interest, associated with high (or perhaps low) trait values. We might ascertain affected subjects (presumably with high trait values) and form two statistics:

1. A regression of quantitative trait on the probability of being type A at a candidate locus.

2. An allele-sharing statistic examining, for example, the average of probabilities of subjects ascertained being of type (A) at a locus, and considering its divergence from the expected value of $\frac{1}{2}$.

We could then combine the two statistics, as discussed here, in hopes of getting a more powerful test. We believe that many more possibilities exist for different types of experiments, and this remains an area for future research.

There are also QTL-mapping statistics in the literature that do not fall into either the IBD sharing or the "trait given IBD" categories but rather draw on both sources of information jointly. Zinn-Justin and Abel (1999) extend the work of Commenges (1994) and Commenges et al. (1994) to introduce IBD information into the weighted pairwise correlation (WPC) statistic. We examined its performance on a limited basis and found the new WPC test nearly as powerful as our composite test for discordant sibs, but much less powerful in applications to concordant siblings than our concordant statistic proposed above (unpublished data). In addition, critical values and standard errors for the WPC must be determined through simulations (Zinn-Justin and Abel 1999), there is currently no multipoint formulation, and the test is not available in widely used packages, as are the components of our discordant composite test. Very recently, both Alcais and Abel (1999) and Dudoit and Speed (2000) have turned the trait-given-IBD distribution around, and devised tests of linkage based on IBD data conditioned on phenotypes, which they argue is the more sensible formulation, since ascertainment is based on phenotype. In both cases, the resultant test statistic is some sort of allele-sharing statistic, with the information on quantitative-trait value entering via a pair-specific weight function.

Acknowledgments

Dr. Feingold's work was partially supported by NIH R01 HG00848 and NIH R01 AG16989-01.

Appendix A

Details of the Observed-Data Log-Likelihood

Consider the complete-data log-likelihood given in (1). Noting that

$$E(\delta_{ij}|\mathbf{Y}_i) = P(Z_i = j|\mathbf{Y}_i)$$

and that

$$E(\log f_\psi(\Phi_i|Z_i = j)|\mathbf{Y}_i) = \log f_\psi(\Phi_i|Z_i = j) \ ,$$

we get the observed-data log-likelihood

$$\begin{aligned} L(\psi,\pi) &= E[L^*(\psi,\pi)|\mathbf{Y}] \\ &= E[L^*(\psi)|\mathbf{Y}] + E[L^*(\pi)|\mathbf{Y}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \log f_\psi(\Phi_i|Z_i = j) \times P(Z_i = j|\mathbf{Y}_i) + \sum_{j=1}^{k} \log \pi_j \sum_{i=1}^{n} P(Z_i = j|\mathbf{Y}_i) \\ &= L_1(\psi,\pi) + L_2(\psi,\pi)\ . \end{aligned} \tag{A1}$$

In general, the probability $P(Z_i = j|\mathbf{Y}_i)$ depends on the locus-specific values of both ψ and π, but when computed under the null hypothesis of no linkage to the candidate locus, it depends on neither. In that case, the maximum likelihood estimator $\hat{\pi}$ is given by

$$\begin{aligned} \hat{\pi}_j &= \frac{\sum_{i=1}^{n} P(Z_i = j|\mathbf{Y}_i)}{\sum_{j=1}^{k}\sum_{i=1}^{n} P(Z_i = j|\mathbf{Y}_i)} \\ &= \frac{1}{n}\sum_{i=1}^{n} P(Z_i = j|\mathbf{Y}_i) \end{aligned} \tag{A2}$$

for $j \in \{1,...,k\}$ (which is analogous to the usual MLE for a multinomial). In order to quantify the dependence of $P(Z_i = j|\mathbf{Y}_i)$ on ψ, we can rewrite it using Bayes' Rule as

$$\begin{aligned} P(Z_i = j|\mathbf{Y}_i) &= P(Z_i = j|\mathbf{M}_i,\Phi_\mathrm{i}) \\ &= P(Z_\mathrm{i} = \mathrm{j}|\mathbf{M}_i) \times \frac{f_\psi(\Phi_\mathrm{i}|Z_i = j)}{f_\psi(\Phi_\mathrm{i}|\mathbf{M}_i)} \\ &= P(Z_\mathrm{i} = \mathrm{j}|\mathbf{M}_i) \times \frac{f_\psi(\Phi_i|Z_i = j)}{\sum_{l=1}^{k} f_\psi(\Phi_i|Z_i = l)P(Z_i = l|\mathbf{M}_i)}\ . \end{aligned} \tag{A3}$$

Equation (A3) facilitates understanding of when the observed-data log-likelihood separates into two distinct functions of ψ and π, respectively. If the marker data $\mathbf{M}_i$ give us perfect information on the inheritance vectors at a locus, we get $P(Z_i = j|\mathbf{Y}_i) = P(Z_i = j|\mathbf{M}_i) \equiv 1$ for some j, so that the dependence on ψ of the second summand in the observed-data log-likelihood disappears. Alternately, under the null hypothesis of no linkage, $f_\psi(\Phi_i|Z_i = j) = f_\psi(\Phi_i|Z_i = l)$ for all values j and l, so that the above expression for $P(Z_i = j|\mathbf{Y}_i)$ reduces to $P(Z_i = j|\mathbf{M}_i)$. In practice, researchers generally replace $\mathrm{P}(Z_i = j|\mathbf{Y}_i)$ by $\mathrm{P}(Z_i = j|\mathbf{M}_i)$, computing the inheritance configuration–sharing probabilities at each locus under the null hypothesis of no linkage to that locus, and yielding an observed-data log-likelihood (3) for which the MLEs $\hat{\psi}$ and $\hat{\pi}$ are asymptotically independent of one another.

A1. *General Maximum-Likelihood Estimates*

We sketch here a general approach for determining the MLEs $\hat{\psi}$ and $\hat{\pi}$ from the observed-data log-likelihood with no assumptions about linkage at the locus considered, or equivalently for $\theta \in [0,\frac{1}{2}]$. For notational convenience, denote the IBD sharing probabilities conditional on the marker data $\{\mathbf{M}_i\}$ and the values of $\pi = \{\pi_1,...,\pi_k\}$ by $\{p_{ij}\}$, where $p_{ij} = P(Z_i = j|\mathbf{M}_i)$. MLEs $(\hat{\psi},\hat{\pi})$ can be computed via the EM algorithm (Dempster et al. 1977). To begin, initialize the parameters to their values under the null hypothesis. In the E-step, first estimate the values of the $\{p_{ij}\}$ using the current value of π, marker data $\{\mathbf{M}_i\}$, and Bayes's Rule. Next, use the $\{p_{ij}\}$, the trait data $\{\Phi_i\}$, and the current value of ψ to compute values $\{\mathrm{P}(Z_i = j|\mathbf{Y}_i)\}$ using (A3). In the M-step, we use the (momentarily) fixed values $\{\mathrm{P}(Z_i = j|\mathbf{Y}_i)\}$ to maximize (A1) with respect to (ψ,π). The value of π maximizing (A1) is given as a function of the values $\{P(Z_i = j|\mathbf{Y}_i)\}$ by (A2), whereas the value of ψ maximizing (A1) will depend on the parametric model specified for modeling the trait given the IBD sharing status. The E-step and M-step are then iterated until convergence is reached. In order for the MLEs to be useful, the data must be well described by the model. As discussed earlier, inducing an ascertainment scheme on the data will tend to make the observed distribution of $(\Phi|Z)$ unwieldy, so the model $f_\psi(\phi|z)$ should be carefully thought out and tested if this course is taken.

Appendix B

Approximations Showing Links to Haseman-Elston and Variance Components

Here we demonstrate a link between our model as written for phenotype given IBD status and more traditional modeling approaches to quantitative-trait data such as Haseman-Elston and variance components. Since implementations of these methods use only marker data to compute IBD sharing, we will assume that $P(Z_i = j|\mathbf{Y}_i) = P(Z_i = j|\mathbf{M}_i)$ throughout. Consider the first term of the complete-data log-likelihood $L^*(\psi)$. For simplicity, consider only the term for the *i*th pedigree, and notice that we can rewrite the complete-data log-likelihood $L^*_i(\psi)$ as

$$\begin{aligned} L^*_i(\psi) &= \sum_{j=1}^{k} \delta_{ij} \log f_\psi(\Phi_i|Z_i = j) \\ &= \sum_{j=1}^{k} 1(Z_i = j) \log f_\psi(\Phi_i|Z_i = j) \\ &= \log f_\psi(\Phi_i|Z_i) \ . \end{aligned}$$

Now, define the expected value of the IBD configuration label Z_i given the data $\mathbf{Y}_i$ by $\mu_i = \mathrm{E}[Z_i|\mathbf{Y}_i]$. Since Z_i takes values in $\{1,...,k\}$, we will have

$$\mu_i = \sum_{j=i}^{k} j \times P(Z_i = j|\mathbf{Y}_i)$$

which will have a meaningful value only if the labels $\{1,...,k\}$ have some tangible interpretation. In Haseman-Elston regression, for instance, the inheritance configuration is incorporated through pairwise IBD sharing values in $\{0,1,2\}$, and μ_i is the expected IBD sharing given the marker data between two relatives.

Similarly define the complete-data log-likelihood $L^*_i(\psi)$, written as a function of IBD configuration Z_i, by the function $G(Z_i) = \log f_\psi(\Phi_i|Z_i)$, where we assume that G is a differentiable function of Z_i. A Taylor expansion of $G(\cdot)$ about μ_i, followed by a conditional expectation given the observed data, yields the first summand of the observed-data log-likelihood, followed by a useful approximation:

$$\begin{aligned} L_i(\psi) &= \mathrm{E}[L^*(\psi)|\mathbf{Y}_i] \\ &= \mathrm{E}[G(Z_i)|\mathbf{Y}_i] \\ &= G(\mu_i) + \mathrm{E}[(Z_i - \mu_i)|\mathbf{Y}_i]\frac{dG}{dZ}(\mu_i) + \frac{1}{2}\mathrm{E}[(Z_i - \mu_i)^2|\mathbf{Y}_i]\frac{d^2G}{dZ^2}(\mu_i) + ..., \end{aligned}$$

since $\mathrm{E}[(Z_i - \mu_i)|\mathbf{Y}_i] = 0$ and $\mathrm{E}[(Z_i - \mu_i)^2|\mathbf{Y}_i] = \mathrm{Var}(Z_i|Y_i)$, we get the approximation

$$\begin{aligned} L_i(\psi) &\approx G(\mu_i) \\ &= \log f_\psi(\Phi_i|Z_i = \mathrm{E}[Z_i|\mathbf{Y}_i]) \\ &= \log f_\psi(\Phi_i|Z_i = \mathrm{E}[Z_i|\mathbf{M}_i]) \qquad \text{(B1)} \end{aligned}$$

where $\mathrm{E}[Z_i|\mathbf{Y}_i] = \mathrm{E}[Z_i|\mathbf{M}_i]$ under the null hypothesis of no linkage to the locus of interest (see Appendix A). The approximation here consists of replacing the random inheritance configuration label Z_i with its expected value conditioned on the observed data, denoted μ_i. This approximation will approach exactness as the marker data become precise (i.e., as inheritance vector information at a locus improves), reflected intuitively by the fact that $\mathrm{Var}(Z_i|\mathbf{Y}_i) \to 0$ as the marker data improve to give us perfect information about Z_i. Our purpose in deriving (B1) is to bridge the gap between our theoretical development and the parametric form of methods currently applied in real linkage studies.

As an example of the above approximation, consider how the Haseman-Elston regression method fits into this framework. For a sibling pair with trait values $X_{i,1}$ and $X_{i,2}$, define the phenotype by $\Phi_i = (X_{i,1} - X_{i,2})^2$. The distribution of Φ_i is very often—practically speaking—unknowable, since it depends on the trait distribution in the population, the joint distribution of the trait between siblings, and the ascertainment process. We rely on the robust nature of linear regression: under the hypothesis of no linkage, the true regression slope is zero under very modest conditions (e.g. $\mathrm{Var}\Phi_i < \infty$ or $\mathrm{E}X^4_{i,j} < \infty$). For purposes of this example, we assume normality of Φ_i. For Z_i, we use the usual sib-pair complete data of 0, 1, or 2 genes shared IBD at the candidate locus.

Our model for the squared trait differences can now be written as $\Phi_i|Z_i \sim \mathrm{N}(\alpha + \beta Z_i, \sigma^2)$ so that the first half of the observed-data log-likelihood is

$$
\begin{aligned}
L(\psi) &= \sum_{i=1}^{n} \sum_{j=0}^{2} \log f_\psi(\Phi_i|Z_i = j) \times \mathrm{P}(Z_i = j|\mathbf{Y}_i) \\
&= -\frac{n}{2}\log 2\pi - n\log\sigma + \sum_{i=1}^{n} \sum_{j=0}^{2} -\frac{1}{2}\left(\frac{\Phi_i - \alpha - \beta j}{\sigma}\right)^2 \times \mathrm{P}(Z_i = j|\mathbf{Y}_i) \ .
\end{aligned}
$$

The approximation (denoted $\hat{L}(\psi)$) using (B1) is

$$
\begin{aligned}
\hat{L}(\psi) &= \sum_{i=1}^{n} \log f_\psi(\Phi_i|Z_i = \mathrm{E}[Z_i|M_i]) \\
&= -\frac{n}{2}\log 2\pi - n\log\sigma + \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{\Phi_i - \alpha - \beta \times \mathrm{E}\{Z_i|M_i\}}{\sigma}\right)^2 \ ,
\end{aligned}
$$

from which the MLE $\hat{\beta}$ of β under $\hat{L}(\psi)$ is just the Haseman-Elston estimate of slope when the squared trait difference Φ is regressed on expected IBD status. This shows that if we formulate the observed data log-likelihood (A1) for the sib-pair problem, then approximate the trait-given-IBD portion using (B1), maximum likelihood estimation will yield the Haseman-Elston slope estimate. Hence, the first summand of our log-likelihood function is “almost” Haseman-Elston in this case.

An analogous approximation is now developed to include variance-components methods under this modeling approach. The basic framework for variance-components analysis of quantitative traits is described in Amos (1994). For a group of m_i relatives' trait values Φ_i, we assume a multinormal distribution for the values, where $\mathrm{E}\Phi_i = \mathrm{X}\beta_\mathrm{i}$ for some known $m_i \times p$ matrix of nonmarker covariates X_i and unknown coefficients β. The covariance structure of Φ_i is given by

$$
\begin{aligned}
&\mathrm{Cov}(\Phi_{i,j}, \Phi_{i,l}) \\
&\qquad = \lambda_{j,l}\sigma_a^2 + \gamma_{j,l}\sigma_d^2 + \zeta_{j,l}\sigma_g^2 + 1(j = l)\sigma_e^2 \ , \qquad \text{(B2)}
\end{aligned}
$$

where

1. the additive component of genetic variance σ_a^2 is weighted by $\lambda_{j,l}$, the expected proportion of genes shared IBD at the candidate locus by relatives j and l, given the family's marker data;
2. the dominance component of genetic variance σ_d^2 is weighted by $\gamma_{j,l}$, the probability that j and l share two genes IBD at the candidate locus;
3. the polygenic component of variance σ_g^2 is weighted by $\zeta_{j,l}$, the expected fraction of genes shared IBD by j and l at an arbitrary locus, based solely on their familial relationship (e.g., $\zeta = \frac{1}{2}$ for full siblings, $\zeta = \frac{1}{4}$ for half siblings, $\zeta = \frac{1}{8}$ for first cousins, etc.). Note in particular that $\zeta_{j,l}$ does not depend on the marker data or the inheritance vector; and
4. each individual trait value contains a component of environmental variance σ_e^2.

The fundamental change under our modeling framework is that we redefine the covariance to depend on a vector function of the ith family's nonfounders' inheritance vector Z_i, so that

$$\begin{aligned}&\mathrm{Cov}(\Phi_{i,j},\Phi_{i,l}|Z_i)\\&\quad = h_{\lambda_{j,l}}(Z_i)\sigma_a^2 + h_{\gamma_{j,l}}(Z_i)\sigma_d^2 + \zeta_{j,l}\sigma_g^2 + 1(j = l)\sigma_e^2\ ,\end{aligned} \tag{B3}$$

where in close analogy to the above definitions,

1. $h_{\lambda_{j,l}}(Z_i) = 0,\ \frac{1}{2}$, or 1, as the number of genes j and l share IBD at the candidate locus has value 0, 1, or 2, respectively.
2. $h_{\gamma_{j,l}}(Z_i) = 1$ if j and l share two genes IBD at the locus, and has value 0 otherwise.

We can then write the first summand of the log-likelihood of the ith family as

$$\begin{aligned}L_i(\psi) &= \sum_{j=1}^{k} \log f_{\beta,\Sigma}(\Phi_i|Z_i = j)\ \times\ \mathrm{P}(Z_i = j|\mathbf{Y}_i)\\&= -\frac{m_i}{2}\log 2\pi + \sum_{j=1}^{k}\left\{-\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(\Phi_i - X_i\beta)'\Sigma_j^{-1}(\Phi_i - X_i\beta)\right\}\ \times\ \mathrm{P}(Z_i = j|\mathbf{Y}_i)\ ,\end{aligned} \tag{B4}$$

where m_i is the number of members in family i, and Σ_j is the covariance matrix of $(\Phi_i|Z_i = j)$, i.e., the covariance of the data given that the inheritance vector is in fact the jth one in some enumeration, and $\mathrm{P}(Z_i = j|M_i)$ is the probability that family i's inheritance vector is of type j, given their trait and marker data $\mathbf{Y}_i$. Such a likelihood will be computationally infeasible for even moderate values of m_i. Our intent, however, is not to propose a practical method, but rather to show how applying the approximation given in (B1) to (B3) and (B4) results in the original covariance model for variance components (B2). Specifically, we plug in $\mathrm{E}\{h_{\lambda_{j,l}}(Z_i)|\mathbf{M}_i\}$ and $\mathrm{E}\{h_{\gamma_{j,l}}(Z_i)|\mathbf{M}_i\}$ for $h_{\lambda_{j,l}}(Z_i)$ and $h_{\gamma_{j,l}}(Z_i)$, respectively. The approximate likelihood $\hat{L}(\psi)$ of n families can then be written as

$$\begin{aligned}\hat{L}(\psi) &= \sum_{i=1}^{n} \log f_{\beta,\Sigma}(\Phi_i|h_{\lambda_{j,l}}(Z_i) = \mathrm{E}\{h_{\lambda_{j,l}}(Z_i)|\mathbf{M}_i\}; h_{\gamma_{j,l}}(Z_i) = \mathrm{E}\{h_{\gamma_{j,l}}(Z_i)|\mathbf{M}_i\})\\&= -\frac{1}{2}(\sum_{i=1}^{n} m_i)\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\left\{\log|\hat{\Sigma}^{(i)}| + (\Phi_i - X_i\beta)'(\hat{\Sigma}^{(i)})^{-1}(\Phi_i - X_i\beta)\right\}\end{aligned}$$

where

$$\hat{\Sigma}_{j,l}^{(i)} = \mathrm{E}\{h_{\lambda_{j,l}}(Z_i)|\mathbf{M}_i\}\sigma_a^2 + \mathrm{E}\{h_{\gamma_{j,l}}(Z_i)|\mathbf{M}_i\}\sigma_d^2 + \zeta_{j,l}\sigma_g^2 + 1(j = l)\sigma_e^2\ ,$$

but since $\mathrm{E}\{h_{\lambda_{j,l}}(Z_i)|\mathbf{M}_i\} = \lambda_{j,l}$ and $\mathrm{E}\{h_{\gamma_{j,l}}(Z_i)|\mathbf{M}_i\} = \gamma_{j,l}$, the covariance expression reduces to

$$\hat{\Sigma}_{j,l}^{(i)} = \lambda_{j,l}\sigma_a^2 + \gamma_{j,l}\sigma_d^2 + \zeta_{j,l}\sigma_g^2 + 1(j = l)\sigma_e^2$$

as in standard variance components.

This links variance components into the class of tests we consider and implies that the parameter estimates and tests of linkage gleaned from variance-components analysis will be approximately independent of the inheritance vector-frequency estimates $\{\hat{\pi}_j\}$ which are combined to produce allele-sharing statistics. Hence, evidence of linkage from variance components might be combined with that from allele-sharing methods.

Appendix C

Extensions of Haseman-Elston

Here we summarize and sketch recent attempts to improve the method of Haseman and Elston (1972). Consider a sample of n sib pairs in which the ith pair have trait values $(X_{i,1},X_{i,2})$. Assume that $E(X_{i,1}) = E(X_{i,2}) = \mu$ and that $\mathrm{Var}(X_{i,1}) = \mathrm{Var}(X_{i,2}) = \sigma_x^2$. Let $\hat{\pi}_i$ denote the expected IBD sharing for the two sibs at a locus, conditional on marker data. Define $\Phi_D^i = (X_{i,1} - X_{i,2})^2$ and $\Phi_S^i = (X_{i,1} + X_{i,2} - 2\mu)^2$. Wright (1997) shows that Φ_D and Φ_S are independent. Drigalenko (1998) derives that for an additive trait $E(\Phi_D|\hat{\pi}) = \alpha_D - \beta\hat{\pi}$ and $E(\Phi_S|\hat{\pi}) = \alpha_S + \beta\hat{\pi}$, where β is the only parameter relevant to linkage analysis. Looking to combine the information from these two independent variables for inference about β, both Drigalenko (1998) and Elston et al. (in press) propose regressing some scaled version of $\frac{1}{4}(\Phi_S - \Phi_D) = (X_1 - \mu)(X_2 - \mu)$ on $\hat{\pi}$ and testing the resultant estimate $\hat{\beta}$ against the null hypothesis $H_0 : \beta = 0$.

Φ_S and Φ_D are not equally informative because $Var(\Phi_D) \equiv \sigma_D^2 = 2(\alpha_D - \beta\hat{\pi})^2$, while $Var(\Phi_S) \equiv \sigma_S^2 = 2(\alpha_S + \beta\hat{\pi})^2$, so that weighted regression is likely to improve power. An estimator $\hat{\beta}$ minimizing the weighted sum of squares

$$L(\alpha_D,\alpha_S,\beta,\sigma_D,\sigma_S)$$
$$= \frac{1}{\sigma_D^2}\sum_{i=1}^{n}(\Phi_D^i - \alpha_D + \beta\hat{\pi}_i)^2 + \frac{1}{\sigma_S^2}\sum_{i=1}^{n}(\Phi_S^i - \alpha_S - \beta\hat{\pi}_i)^2$$

can be shown by simulation to be more powerful than both the old and new Haseman-Elston tests and about as powerful under many models as the likelihood-ratio test statistic based on variance-components analysis of sib pairs. This approach will be discussed in detail in a future publication.

References

de Andrade M, Theil TJ, Yu L, Amos CI (1997) Assessing linkage on chromosome 5 using components of variance approach: univariate versus multivariate. Genet Epidemiol 14:773–778

Alcais A, Abel L (1999) Maximum-likelihood-binomial method for genetic model-free linkage analysis of quantitative traits in sibships. Genet Epidemiol 17:102–117

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, Krushkal J, Thiel TJ, Young A, Zhu DK, Boerwinkle E, de Andrade M (1997) Comparison of model-free linkage mapping strategies for the study of a complex trait. Genet Epidemiol 14:743–748

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85–97

Commenges D (1994) Robust genetic linkage analysis based on a score test of homogeneity: The weighted pairwise correlation statistic. Genet Epidemiol 11:189–200

Commenges D, Olson J, Wijsman E (1994) The weighted rank pairwise correlation statistic for linkage analysis: Simulation Study and application to Alzheimer's disease. Genet Epidemiol 11:201–212

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc B 39:1–38

Donnelly KP (1983) The probability that related individuals share some section of the genome identical by descent. Theor Popul Biol 23:34–63

Drigalenko E (1998) How sib pairs reveal linkage. Am J Hum Genet 63:1242–1245

Dudoit S, Speed TP (1999) A score test for linkage using identity by descent data from sibships. Ann Stat 27:943–986

——— (2000) A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data on sib-pairs. Biostatistics 1:1–26

Duggirala R, Blangero J, Almasy L, Dyer TD, Williams KL, Leach RJ, O'Connell P, et al (1999) Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. Am J Hum Genet 64:1127–1140

Elston RC, Buxbaum S, Jacobs KB, Olson JM. Haseman and Elston revisited. Genet Epidemiol (in press). Download at http://darwin.cwru.edu/he2/

Fisher SE, Marlow AJ, Lamb J, Maestrini E, Williams DF, Richardson AJ, Weeks DE, et al (1999) A quantitative-trait locus on chromosome 6p influences different aspects of developmental dyslexia. Am J Hum Genet 64:146–156

Fulker DW, Cherny SS (1996) An improved multipoint sibpair analysis of quantitative traits. Behav Genet 26:527–532

Ghosh S, Watanabe RM, Hauser ER, Valle T, Magnuson VL, Erdos MR, Langefeld CD, et al (1999) Type 2 diabetes: Evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. Proc Natl Acad Sci 96:2198–2203

Gu C, Todorov AA, Rao DC (1996) Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of QTLs. Genet Epidemiol 13:513–533

——— (1997) Genome screening using extremely discordant and extremely concordant sib pairs. Genet Epidemiol 14: 791–796

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

Knapp M (1998) Evaluation of a restricted likelihood ratio test for mapping quantitative trait loci with extreme discordant sib pairs. Ann Hum Genet 62:75–87

Kruglyak L, Daley MJ, Reeve-Daly MP, Lander ES (1996) Parametric and non-parametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Kruse R, Seuchter SA, Baur MP, Knapp M (1997) The "possible triangle" test for extreme discordant sib pairs. Genet Epidemiol 14:833–838

MacCleur JW, Blangero J, Dyer TD, Speer MC (1997) GAW10: simulated family data for a common oligogenic disease with quantitative risk factors. Genet Epidemiol 14:737–742

McPeek MS (1999) Optimal allele-sharing statistics for genetic mapping using affected relatives. Genet Epidemiol 16: 225–249

Nicolaou M, Premkumar S, DeStefano AL, Farrer LA, Cupples LA (1999) Power of concordant versus discordant sib pairs at different penetrance levels. Genet Epidemiol 17: S679–S684

Pugh EW, Jaquish CE, Sorant AJM, Doetsch JP, Bailey-Wilson JE, Wilson AF (1997) Comparison of sib-pair and variance components methods for genomic screening. Genet Epidemiol 14:867–872

Risch N, Zhang H (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268: 1584–1589

——— (1996) Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. Am J Hum Genet 58:836–843

Rochberg DR, Rochberg N, Hampe CL, Suarez BK (1997) The utility of deviant sib pairs in the detection of linkage. Genet Epidemiol 14:873–878

Rogus JJ, Harrington DP, Jorgenson E, Xu X (1997) Effectiveness of extreme discordant sib pairs to detect oligogenic disease loci. Genet Epidemiol 14:879–884

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

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

Xu X, Rogus JJ, Terwedow HA, Yang J, Wang Z, Chen C, Niu T, et al (1999) An extreme-sib-pair genome scan for genes regulating blood pressure. Am J Hum Genet 64: 1694–1701

Zinn-Justin A, Abel L (1999) Introduction of the IBD information into the weighted pairwise correlation method for linkage analysis. Genet Epidemiol 17:35–50