Mapping quantitative traits with random and with ascertained sibships

Jie Peng and D. Siegmund*

Department of Statistics, Stanford University, Stanford, CA 94305

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected on April 30, 2002.

Contributed by D. Siegmund, March 10, 2004

Use of a robust score statistic based on a variance components model to map quantitative trait loci in randomly sampled pedigrees is reviewed. Sibships ascertained through a single proband are discussed. Under a standard assumption of multivariate normality, two suggested methods of ascertainment correction are shown to be asymptotically equivalent when the number of sibships is large.

A seminal contribution to mapping quantitative trait loci (QTL) in humans is the regression method of Haseman and Elston (1). In the past decade this method has, to a considerable degree, been superceded by variance component methods (e.g., refs. 2–6), which are typically more flexible with regard to pedigree structure and more powerful (e.g., refs. 7 and 8). A recent regression-based contribution that contains many of the positive features of variance component methods is provided by Sham *et al.* (9). See Feingold (10) for a review of and additional references to a rapidly expanding literature.

Although the basic theory associated with these methods presupposes that pedigrees are randomly sampled, in practice, they are often ascertained through one or more probands having particular phenotypic values, e.g., a proband who has an extreme phenotype, perhaps by virtue of being affected by a disease for which the quantitative trait is a diagnostic marker.

This article reviews recent methodological developments in QTL mapping derived under the assumption of random sampling and gives a more detailed analysis of one simple ascertainment method. We begin by reviewing some basic theory, which is closely related to and combines features of both regression and variance components methods. In particular, we review the argument of Tang and Siegmund (11) that a particular parameterization and systematic use of the large-sample statistical theory of score statistics allows one to compute explicitly what would otherwise be very complicated expressions, and, consequently, to understand results that previously were inferred from extensive numerical simulations. *Comparison with Regression Methods: Miscellaneous Remarks* contains a brief comparison of our method with regression-based methods along with discussion of the underlying assumptions and ways to deal with violations of those assumptions. In *Ascertainment* we build on the results of first sections to give an analysis of single ascertainment, in particular, a demonstration that two apparently different methods of ascertainment correction (2, 12) are asymptotically equally powerful when the number of pedigrees is large. The case of an arbitrary number of probands and more complete numerical results will be discussed elsewhere. The final section contains a discussion of the implications and limitations of our ascertainment corrections.

Description of the Model

We assume Hardy–Weinberg and linkage equilibrium throughout. This assumption means, in particular, that, for both markers and QTL, haplotypes within the same locus and genotypes among different loci are stochastically independent. Our basic model goes back to the classic article by Fisher (13) for the case of diallelic genes; the general case is discussed by Kempthorne (14). We assume a QTL exists at the genomic location τ. The phenotypic value Y is assumed to be given by

$$Y = \mu + \alpha_x + \alpha_y + \delta_{x,y} + e. \qquad [1]$$

The mean value μ can also accommodate covariates in the form of a linear model with minor changes to what follows. The parameter $\alpha_a = \alpha_a(\tau)$ denotes the additive genetic effect of allele a at locus τ; $\delta_{a,b}$ denotes the dominance deviation of alleles a and b. A subscript x denotes the allele contributed by the mother, whereas a subscript y refers to the father. By standard analysis of variance arguments, we may assume that $E\alpha_x = E\alpha_y = E(e) = E[\delta_{x,y}|x] = E[\delta_{x,y}|y] = 0$. Since by the assumption of Hardy–Weinberg equilibrium x and y are independent (unless the parents are inbred), the different genetic effects in Eq. **1** are uncorrelated. We assume, in addition, that e is uncorrelated with the explicitly modeled genetic effects. The phenotypic variance is $\sigma_Y^2 = E[(Y - \mu)^2]$. The variances of the additive and dominance effects associated with the QTL at τ are by definition $\sigma_A^2 = 2E\alpha_x^2$ and $\sigma_D^2 = E\delta_{x,y}^2$. Implicitly, we expect that several QTL may occur, which may interact. (An explicit model is given below). For this article we assume that other QTL lie on other chromosomes and are in linkage equilibrium with the QTL at τ. Then, their contribution to the phenotype Y can be assumed to be a part of the residual term e. With the notation $\sigma_e^2 = \mathrm{Var}(e)$, it follows that $\sigma_Y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_e^2$.

Now consider a pair of siblings satisfying the model (Eq. **1**). Recall that at any locus two relatives share alleles identical by descent if they inherit the same alleles from a common ancestor. Two siblings can share 2, 1, or 0 alleles identical by descent depending on whether they inherit the same alleles from both mother and father, from one but not both, or from neither. Let $\nu = \nu(\tau)$ denote the number of alleles identical by descent at τ. Letting Y_i denote the phenotypic value of the ith sibling ($i = 1, 2$), we have (refs. 13 and 14):

$$\mathrm{Cov}[Y_1, Y_2|\nu] = \sigma_e^2 r + \sigma_A^2 \nu/2 + \sigma_D^2 1_{\{\nu=2\}}, \qquad [2]$$

where $r = \mathrm{corr}(e_1, e_2)$ accounts for the correlation between sibs that arises from other QTL and from a shared environment.

Taking the expectation of Eq. **2**, we find the unconditional covariance. Then we can rewrite Eq. **2** in the form

$$\mathrm{Cov}[Y_1, Y_2|\nu] = \mathrm{Cov}(Y_1, Y_2) + [(\sigma_A^2 + \sigma_D^2)/2](\nu - 1) - (\sigma_D^2/2) \cdot [1\{\nu = 1\} - 1/2].$$

In this equation the terms involving ν have mean 0 and are uncorrelated. In what follows it will be convenient to introduce new parameters $\rho_\nu = \mathrm{Cov}[Y_1, Y_2|\nu]/\sigma_Y^2$, $\rho = \mathrm{Cov}(Y_1, Y_2)/\sigma_Y^2$,

Abbreviation: QTL, quantitative trait loci.

See accompanying Biography on page 7843.

*To whom correspondence should be addressed. E-mail: dos@stat.stanford.edu.

$\alpha_0 = (\sigma_A^2 + \sigma_D^2)/2$, and $\delta_0 = \sigma_D^2/2$, and rewrite the preceding equation in the form

$$\rho_\nu = \rho + \sigma_Y^{-2}\{\alpha_0(\nu - 1) + \delta_0[1/2 - 1\{\nu = 1\}]\} \qquad [3]$$

(compare with Eq. **4**).

Since the QTL location, τ, is unknown, we will be interested in marker loci t distributed throughout the genome and the process $\nu(t)$ for a sib pair considered as a stochastic process in t. For markers t_1 and t_2 on different chromosomes, $\nu(t_1)$ and $\nu(t_2)$ are stochastically independent. For markers on the same chromosome $\mathrm{Cov}[\nu(t_1), \nu(t_2)] = 2^{-1}[1 - 2\phi]$, where ϕ is a function of the recombination frequency. When recombination follows the Haldane model of no interference, $1 - 2\phi = \exp(-4|t_1 - t_2|)$, where the marker location t_i denotes genetic distance in morgans (M) from a designated end of the chromosome.

In this article we assume that $\nu(t)$ is observable. This assumption is often not satisfied and the process of estimating the value of $\nu(t)$ from marker data is quite complex [e.g., Kruglyak *et al.* (15)]. See Teng and Siegmund (16) for a theoretical analysis of the amount of information lost in this process.

Score Statistics

Suppose we have a sample of N sibships, each of size s. We index sibs within a sibship by i and j and sibships by $n = 1, \ldots N$. The subscript n is often suppressed in our notation. Let $Y = Y_n$ denote the vector of phenotypes in the nth sibship. Let $\nu_{ij}(t)$ denote the number of alleles shared identical by descent at the marker locus t by the ith and jth sibs in the nth sibship. Let A_ν denote the $s \times s$ matrix with entries $\nu_{ij} - 1$ for $i \neq j$ and zeroes along the diagonal, and let D_ν denote a similar matrix with off diagonal elements $(1/2 - 1\{\nu_{ij} = 1\})$. Let $\Sigma_\nu = E[(Y - \mu)(Y - \mu)'|A_\nu)$ and $\Sigma = E[(Y - \mu)(Y - \mu)']$, so from Eq. **3**, we have

$$\Sigma_\nu = \Sigma + \alpha_0 A_\nu + \delta_0 D_\nu. \qquad [4]$$

The critical assumption of components of variance linkage analysis is that conditional on A_ν, the random variable Y has a multivariate normal distribution. This assumption has mathematical convenience to facilitate the following computations. It cannot be expected to be exactly true even, or perhaps especially, in the case that the QTL are diallelic, so it is important to check (as we discuss below) that the statistical consequences of this assumption are reasonable.

Under the normality assumption, the log likelihood for the QTL at τ is $\ell = \ell(\tau, \alpha_0, \delta_0, \rho)$ given by

$$\ell = -2^{-1}\sum_{n=1}^{N}\{\log|\Sigma_\nu| + \mathrm{tr}\Sigma_\nu^{-1}(Y - \mu)(Y - \mu)'\}, \qquad [5]$$

where $\nu = \nu(\tau)$.

Using the identities $\partial \log|G|/\partial x = \mathrm{tr}(G^{-1}\partial G/\partial x$ and $\partial G^{-1}/\partial x = -G^{-1}\partial G/\partial x G^{-1})$, which are valid for any differentiable nonsingular matrix function, we obtain the score equations:

$$\ell_\alpha = 2^{-1}\sum_n \{-\mathrm{tr}(\Sigma_\nu^{-1}A_\nu) + \mathrm{tr}(\Sigma_\nu^{-1}A_\nu\Sigma_\nu^{-1}(Y - \mu)(Y - \mu)')\} \qquad [6]$$

and

$$\ell_\rho = 2^{-1}\sum_n \{-\mathrm{tr}(\Sigma_\nu^{-1}B) + \mathrm{tr}(\Sigma_\nu^{-1}B\Sigma_\nu^{-1}(Y - \mu)(Y - \mu)')\},$$

where $B = \partial\Sigma_\nu/\partial\rho = \mathbf{1}\mathbf{1}' - I$. We omit the similar expression for ℓ_δ. The Fisher information matrix can be computed as the expected value of

$$-\ell_{\alpha\alpha} = \sum_n \{-2^{-1}\mathrm{tr}(\Sigma_\nu^{-1}A_\nu\Sigma_\nu^{-1}A_\nu) + \mathrm{tr}(\Sigma_\nu^{-1}A_\nu\Sigma_\nu^{-1}A_\nu\Sigma_\nu^{-1}(Y - \mu)(Y - \mu)')\} \qquad [7]$$

and similar expressions for $\ell_{\rho\rho}$, $\ell_{\alpha\rho}$, $\ell_{\delta\delta}$, etc. Under the null hypothesis that α_0 (hence, also δ_0) $= 0$, the scores ℓ_α and ℓ_δ for the parameters of interest are linear functions of A_ν and D_ν, which have mean values equal to 0, so they are uncorrelated with the scores ℓ_ρ, $\ell_{\sigma_Y^2}$, and ℓ_μ for the segregation parameters, which depend only on phenotypic data (when $\alpha_0 = 0$).

The score statistic at a given marker location t to test the hypothesis that $\alpha_0 = 0$ is

$$Z_t = \ell_\alpha(t)/I_{\alpha\alpha}^{1/2}. \qquad [8]$$

Here, $I_{\alpha\alpha}$ is the appropriate entry from the Fisher information matrix (i.e., the expected value of Eq. **7**); both numerator and denominator are evaluated with $\alpha_0 = \delta_0 = 0$ and with the nuisance parameters estimated by their maximum likelihood estimates under the condition that $\alpha_0 = \delta_0 = 0$. A second statistic involving ℓ_δ, which is asymptotically uncorrelated with Eq. **8**, can be defined similarly. For ease of exposition we temporarily ignore this second statistic.

It may be shown [Tang and Siegmund (11)] that at a marker t linked to the trait,

$$E(Z_t) \approx N^{1/2}\xi(1 - 2\phi), \qquad [9]$$

where ϕ is as given above and

$$\xi^2 = \frac{\alpha_0^2}{2\sigma_Y^4}\binom{s}{2}\frac{\{[1 + (s - 2)\rho]^2 + \rho^2\}}{\{(1 - \rho)[1 + (s - 1)\rho]\}^2}. \qquad [10]$$

The denominator of the score statistic (Eq. **8**) is very sensitive to the assumption of normality. This sensitivity can be mitigated somewhat by considering the conditional distribution of the numerator given the phenotypic observations, $Y_1, \ldots, Y_N$, which, under the hypothesis $\alpha_0 = \delta_0 = 0$, is approximately a normal distribution with mean 0 and standard deviation $\{E_0[\ell_\alpha^2|Y_1, \ldots, Y_N]\}^{1/2}$ in large samples. Tang and Siegmund (11) have calculated this quantity, which, when used to standardize $\ell_\alpha(t)$ instead of $I_{\alpha\alpha}^{1/2}$, leads to a statistic that is asymptotically normally distributed whether the normality hypothesis is satisfied or not, and it still has roughly the noncentrality parameter (Eq. **10**) if the normality assumption is approximately true.

In the special case of sib pairs, this robust score statistic takes a relatively simple form. One can rewrite Eq. **6** in terms of the uncorrelated variables $D = (Y_1 - Y_2)$ and $S = Y_1 + Y_2 - 2\mu$, to obtain

$$\ell_\alpha(t) = \sum_n [\nu(t) - 1]C_n,$$

where

$$C_n = \rho_\nu/(1 - \rho_\nu^2) + S_n^2/2\sigma_Y^2(1 + \rho_\nu)^2 - D_n^2/2\sigma_Y^2(1 - \rho_\nu)^2.$$

We set $\alpha_0 = \delta_0 = 0$ and replace the segregation parameters μ, σ_Y^2, and ρ by their maximum likelihood estimators under the condition $\alpha_0 = \delta_0 = 0$ to obtain, say, $\hat{C}_n$. The robust score statistic at the marker locus t is

$$\sum_n [\nu(t) - 1]\hat{C}_n \Big/ \left[\sum_n \hat{C}_n^2/2\right]^{1/2}. \qquad [11]$$

Its asymptotic expectation is $N^{1/2}$ times

$$\xi = \frac{\alpha_0}{\sigma_Y^2} \frac{(1+\rho^2)/[(1-\rho^2)^2]}{[2E_0C_n^2]^{1/2}}.$$

In the normal case this reduces to Eq. **10** with $s = 2$.

Genome Scans

Since we do not know the location of the QTL τ, we scan the genome using

$$Z_{\max} = \max_t Z_t,$$

where the max is taken over all marker loci t. To use this statistic, one must establish a detection threshold, $z_{\max}$, which must be large enough to avoid false-positive errors and small enough to allow detection of true signals. For data involving a large number of pedigrees, so the statistic (Eq. **8**) or the robust alternative suggested above is approximately normally distributed, for an idealized genome scan with markers equally spaced at a distance Δ cM in a genome containing c chromosomes of total length L (cM), the following approximation to the genome-wide false-positive rate is given by Feingold *et al.* (17). Writing P_0 to denote probability under the hypothesis that $\alpha_0 = \delta_0 = 0$, we have

$$P_0\left\{\max_i Z_{i\Delta} \geq z\right\} \approx 1 - \exp\{-c[1-\Phi(z)] - L\beta z\varphi(z)h[z(2\beta\Delta)^{1/2}]\}, \quad [12]$$

where φ and Φ are the standard normal density and distribution functions, respectively, and h is the special function discussed by Siegmund (ref. 18, p. 82). The function h can be computed numerically, but the arguments given on pages 210–211 suggest the simple approximation

$$h(2x) \approx \frac{x^{-1}[\Phi(x) - 1/2]}{x\Phi(x) + \varphi(x)}.$$

Some numerical exploration shows that this is a surprisingly good approximation for all $x > 0$. For a comprehensive discussion of genome-wide significance thresholds in linkage analysis see Lander and Kruglyak (19).

As a numerical example, for sib pairs, a 22-chromosome 3,300-cM human genome and markers equally spaced at 5 cM, a threshold of approximately $z_{\max} = 3.73$ produces the conventional 0.05 false-positive error rate. Similar approximations for the power (17) show that a noncentrality parameter of $N^{1/2}\xi = 5$ produces power of $\approx$0.91 to detect a QTL located at a marker and 0.87 when the QTL is midway between markers (compare Eqs. **9** and **10**). The adequacy of the approximation (Eq. **12**) and modifications to deal with different formulations are discussed below.

Comparison with Regression Methods: Miscellaneous Remarks

The preceding argument is in the spirit of variance components, which are often contrasted with "regression-based" methods. The conventional wisdom is that variance components are more flexible in dealing with large pedigrees and more efficient when the normality assumption is approximately satisfied but less robust against violations of the assumption of normality (e.g., refs. 9 and 10). However, the efficient score, ℓ_α, derived from a variance components model is of the form of a covariance, so if it is standardized by a nonparametric estimator of its standard deviation, as suggested above, one obtains a regression-like statistic (compare Eq. **11**) that is robust against nonnormality under the null hypothesis of no linkage. In fact, the original Haseman–Elston regression statistic for sib pairs can be deduced by a similar line of reasoning by starting from the likelihood function for D alone and ignoring S. The "new" Haseman–Elston statistic (20) can be derived by starting with the likelihood function for $S^2 - D^2$.

It is straightforward to show that when the phenotypes are close to normally distributed the asymptotic squared noncentrality parameter (per sib pair) for the classical Haseman–Elston statistic is

$$\frac{\alpha_0^2}{2\sigma_Y^4} \frac{1}{2(1-\rho)^2},$$

whereas that of the new Haseman–Elston statistic is

$$\frac{\alpha_0^2}{2\sigma_Y^4} \frac{1}{1+\rho^2}.$$

By comparing these statistics with Eq. **10** with $s = 2$, one sees that under the normality assumption (Eq. **8**) has greater asymptotic power, sometimes much greater, than either of the Haseman–Elston statistics, the first of which has comparable power when ρ is large, whereas the second has comparable power when ρ is small.

The Haseman–Elston approach is inefficient because it reduces data that is fundamentally two-dimensional (when $s = 2$) to one dimension. A multivariate regression-based method was introduced recently by Sham *et al.* (9), and it appears to be essentially equivalent to the robust variance components method described above. In particular, Sham *et al.* show that for their method the asymptotic noncentrality parameter of a sibship of size s is also given by Eq. **10**.

A possible advantage of the regression approach is that there is some flexibility in the assumed covariance of the dependent variables, hence, in the weights of the resulting generalized least-squares estimators. Sham *et al.* (9) choose a covariance function that would be optimal under an assumption of multivariate normality. Xu *et al.* (21) suggest a different choice, which they have developed only for sib pairs but which, according to Cuenco *et al.* (22), may have some advantages when trait distributions are far from normal.

Remarks

(*i*) In our analysis the primary role of the normality assumption is to suggest the form of the statistic given in Eq. **6**, which as noted above can be regarded as a covariance between a function of phenotypes and identity by descent counts. An alternative to the normality assumption that is equally tractable is the multivariate t distribution [cf. Lange *et al.* (23)]. This assumption leads to a similar statistic, but, because of the heavier tails of the t distribution, it is more robust to outliers in the data. However, it cannot avoid problems that arise from modeling the complexities of multivariate dependence by multivariate distributions that measure dependence only by pairwise correlations.

(*ii*) In the preceding argument we assumed completely informative markers to simplify the analysis and made the working assumption that the QTL τ is one of the markers. If either of these assumptions fails to be true, the likelihood function involves a mixture based on the conditional distribution of $\nu(\tau)$ given the marker data, say M, in the nth family. A convenient representation for the likelihood function is $E_0[\exp(\ell(\tau, \alpha_0, \delta_0, \rho) - \ell(0, 0, \rho))|M, Y]$, where M denotes marker data, ℓ is given by Eq. **5**, and E_0 denotes expectation under the hypothesis that $\alpha_0 = \delta_0 = 0$. When this hypothesis holds, one sees from Eq. **6** that ℓ_α is linear in A_ν. Hence, the numerator of the score statistic for partially informative markers is the same as for fully informative markers, but A_ν is replaced by its conditional expectation $A_{\hat{\nu}} = E_0[A_\nu|M]$. The likelihood ratio statistic is nonlinear in the A_ν.

Hence, it requires a more complicated calculation, although it has been observed in Monte Carlo studies (e.g., ref. 4) that simply replacing A_ν by $E_0[A_\nu|M]$ can produce excellent results. Since the score and likelihood ratio statistics are asymptotically equivalent when α_0 and δ_0 are small, the preceding observation about the score statistic provides a theoretical basis for understanding these Monte Carlo results.

The denominator of the score statistic must also be modified to account for partially informative markers. For example, for sib pairs the 1/2 in the denominator of Eq. **11** arises there as the value of $E_0[(\nu(t) - 1)^2]$, which can be estimated by, for example, $N^{-1} \sum_1^N (\hat{\nu}(t) - 1)^2$. For other estimators and a comparative evaluation see Cuenco *et al.* (22).

The effect of partially informative markers is to increase the autocorrelation of the score statistic and hence to make the approximation given in Eq. **12** somewhat conservative. Although this issue has not received a satisfactory theoretical analysis, there is numerical evidence that the effect is relatively modest (cf. ref. 16).

(*iii*) Other reasons exist that the approximation Eq. **12** may fail to be adequate. The most important is that the distribution of Z_t can fail to be approximately normal. This distribution can be skewed if large sibships or pedigrees containing more distant relatives than siblings are involved, and it can have excess kurtosis if the phenotypic distributions do. Tang and Siegmund (11) suggest a modification of Eq. **12** that accounts for skewness. This approximation can also be adapted to account for kurtosis. The parameters for skewness and kurtosis should be determined from the conditional distribution of Z_t given the phenotypes and will involve the empirical distribution of the phenotypes.

Although Eq. **10** suggests that large sibships may be substantially more powerful than small sibships, a larger threshold is also required because of skewness in the distribution of Eq. **8**. Tang and Siegmund (11) show that after adjusting for the larger threshold the power of large sibships turns out to be considerable, although it is not as great as it would appear from the noncentrality parameter alone.

(*iv*) If it is thought that dominance may play an important role, one can also consider a second degree of freedom ℓ_δ, which is uncorrelated with ℓ_α when $\alpha_0 = \delta_0 = 0$. At a QTL τ it has a noncentrality parameter proportional to $\delta_0 = \sigma_D^2/2$. (The constant of proportionality is the same as in Eq. **10**, except that 2 is replaced by 4 in the denominator.) However, since $\alpha_0 = (\sigma_A^2 + \sigma_D^2)/2$ involves the dominance variance and exceeds $\delta_0 = \sigma_D^2/2$, it turns out that the second degree of freedom rarely adds substantially to the power to detect linkage. See ref. 24 for the modification to Eq. **12** required by the two-dimensional statistic and the constraint $0 \le \delta_0 \le \alpha_0$.

(*v*) The model (Eq. **1**) is very flexible in many respects. For example, it can accommodate multivariate phenotypes, although this accommodation leads to multivariate statistics and hence requires a higher detection threshold to maintain the same false-positive error rate. If one uses for two phenotypes a 3 df statistic involving additive effects on each of the phenotypes and the correlation of these effects, an extension of the method of Dupuis and Siegmund (24) allows one to determine approximate significance thresholds and power. In comparison with the threshold $z_{\max} = 3.73$ and noncentrality of $N^{1/2}\xi \approx 5$ for ≈90% power discussed above for a single trait, the corresponding threshold and noncentrality parameter would increase to $z_{\max} = 4.50$ and $N^{1/2}\xi \approx 5.50$. Such an increase in noncentrality would be greatest when QTL for each trait are tightly linked or even identical because of pleiotropy. A detailed numerical study is required to determine more precisely the conditions under which use of multivariate phenotypes is advantageous. Wang (25) contains a related discussion.

(*vi*) The model (Eq. **1**) can also be expanded to include multiple, possibly interacting, QTL. Assuming for simplicity that no dominance exists, then for two QTL at unlinked loci τ and $\tilde{\tau}$, the phenotype Y is given by

$$Y = \mu + \alpha_x + \alpha_y + \tilde{\alpha}_{\tilde{x}} + \tilde{\alpha}_{\tilde{y}} + \gamma_{x,\tilde{x}} + \gamma_{x,\tilde{y}} + \gamma_{y,\tilde{x}} + \gamma_{y,\tilde{y}} + e.$$

Here α_a denotes the additive effect of allele a at locus τ, $\gamma_{a,\tilde{a}}$ denotes the additive–additive interaction of alleles a at τ and $\tilde{a}$ at $\tilde{\tau}$, etc. As before, σ_A^2 is twice the variance of the additive effect α_x, whereas $\sigma_{A\tilde{A}}^2 = 4E[\gamma_{x,\tilde{x}}^2]$ is the additive–additive interaction variance. A perhaps surprising feature of this model is that the essential ingredient of the noncentrality parameter of ℓ_α is now $\alpha_0 = \sigma_A^2/2 + \sigma_{A\tilde{A}}^2/4$, i.e., a fraction of the interaction variance involving the QTL at τ and $\tilde{\tau}$ enters into the noncentrality of the score statistic that tests for an additive effect at τ only. The score for the additive–additive interaction effect has expectation proportional to $\gamma_0 = \sigma_{A\tilde{A}}^2/4$. Its squared noncentrality equals 1/2 of Eq. **10** with α_0 replaced by γ_0, so much of the effect of the interaction variance component appears in the noncentrality parameter of the statistic to test for a main effect. This fact is similar to the phenomenon in Remark *iv* regarding dominance and is quite different from the situation in experimental genetics, where in a backcross or intercross the noncentrality parameters of statistics that test for main effects are unaffected by interactions with unlinked QTL.

(*vii*) Some of the calculations reported above about the asymptotic noncentrality parameter of the robust score statistic, when the normality assumption is violated, are based on the fact that the expected value of ℓ_α is asymptotically the same as the value one obtains when the normality assumption holds. At first glance, this equality seems almost obvious, since for known nuisance parameters evaluation of $E[\ell_\alpha]$ depends only on the validity of Eq. **4**, which in turn depends only on the basic genetic model, not the normality assumption. However, the segregation parameters μ, σ_Y^2, and ρ must be estimated, so one must show that this effect, which, under the normality assumption is negligible as a consequence of the orthogonality of the segregation and linkage parameters, is also negligible without that assumption. This effect can be demonstrated by a lengthy Taylor series approximation coupled with the observation that in the term contributed by the nth pedigree the nuisance parameters can be estimated almost equally well by the phenotypic variables from the other $N - 1$ pedigrees, which then would give an estimate that is independent of the data in the nth pedigree. We omit the details.

Ascertainment

When pedigrees are ascertained by random sampling, the nuisance parameters μ, σ_Y^2, and ρ are easily estimated. In many cases, however, pedigrees are ascertained through the phenotypes of one or more probands, and phenotypes are determined only for ascertained pedigrees. Here we consider the simplest possible situation, where each sibship contains one proband, with phenotypic value Y_1; and we are particularly interested in the case where ascertainment is based on a threshold T, so a sibship is ascertained if the proband's phenotype satisfies $Y_1 \ge T$. As we observe below, the efficient score, ℓ_α, has the same form as in the case of random ascertainment; but the estimators of segregation parameters involve an ascertainment correction.

Single ascertainment, as described in the preceding paragraph, has been studied by Elston and Sobel (12), who suggest that one correct for ascertainment by conditioning on the event that a pedigree is ascertained, and by Hopper and Mathews (2), who suggest conditioning on the phenotypic value of the proband (cf. also ref. 26). Using simulation, Andrade and Amos (27) have compared these suggestions and have found that the two methods are comparable. Below we show that, in fact, they are asymptotically equivalent when the number of sibships is large, so the results obtained by simulation are exactly as expected.

Ascertainment based on an arbitrary number of probands and more detailed numerical results when the ascertainment rule is not so easily specified will be discussed in a future paper.

The phenotypic vector Y can be partitioned into $(Y_1, Y^{(2)})$, where Y_1 is the phenotype of the ascertained sibling. We begin by considering a conditional analysis of $Y^{(2)}$ give the value Y_1. For notational simplicity we assume $\sigma_Y^2 = 1$. Because the efficient score for σ_Y^2 turns out to be uncorrelated with the efficient score for α_0, this has no effect on the asymptotic theory that follows. Let $\mu_\nu = E(Y^{(2)}|Y_1, A_\nu)$. Assume for simplicity that there is no dominance, i.e., δ_0 in Eq. **4** equals zero. Then $\mu_\nu = \mu\mathbf{1} + (Y_1 - \mu)(\rho\mathbf{1} + \alpha_0 a_\nu)$, where $\mathbf{1}$ is an $s - 1$ dimensional vector with 1 at each entry, ρ and α_0 have the same meaning as above, and $a_\nu' = (\nu_{12} - 1, \ldots, \nu_{1s} - 1)$. Also let

$$\begin{aligned}\Sigma_{2,\nu} &= \mathrm{Cov}(Y^{(2)}|Y_1, A_\nu)\\ &= \mathrm{Cov}(Y^{(2)}) + \alpha_0 A_\nu - (\rho\mathbf{1} + \alpha_0 a_\nu)(\rho\mathbf{1} + \alpha_0 a_\nu)'.\end{aligned}$$

The conditional log likelihood given Y_1, A_ν is exactly of the form of Eq. **5**, but with Y replaced by $Y^{(2)}$, μ replaced by μ_ν, Σ_ν replaced by $\Sigma_{2,\nu}$, and the sum is over ascertained sibships.

It is readily verified that the derivative with respect to α_0 of $\Sigma_{2,\nu}$ *is* $\dot\Sigma_{2,\nu} = A_{2,\nu} - \rho B_\nu$, where $B_\nu = a_\nu\mathbf{1}' + \mathbf{1}a_\nu'$. The efficient score for α_0 evaluated at $\alpha_0 = 0$ is

$$\begin{aligned}\ell_\alpha = \sum_n \{&[-\mathrm{tr}[\Sigma_2^{-1}\dot\Sigma_{2,\nu}]/2 + (Y_1 - \mu)a_\nu'\Sigma_2^{-1}(Y^{(2)} - \mu^{(2)})\\ &+ (Y^{(2)} - \mu^{(2)})'\Sigma_2^{-1}\Sigma_{2,\nu}\Sigma_2^{-1}(Y^{(2)} - \mu^{(2)})/2\},\end{aligned} \qquad [13]$$

where $\mu^{(2)}$ is μ_ν evaluated at $\alpha_0 = 0$ and $\Sigma_2 = \mathrm{Cov}(Y^{(2)}|Y_1)$. Expressions can also be obtained for ℓ_ρ and ℓ_μ, which when $\alpha_0 = 0$ do not depend on ν, hence are conditionally, given Y_1, uncorrelated with ℓ_α.

From the second derivative, $\ell_{\alpha\alpha}$, one finds that when $\alpha_0 = 0$

$$\begin{aligned}E_0(-\ell_{\alpha\alpha}|A_\nu, Y_1) = \sum_n \{&\mathrm{tr}[\Sigma_2^{-1}\dot\Sigma_{2,\nu}\Sigma_2^{-1}\dot\Sigma_{2,\nu}]/2\\ &+ (Y_1 - \mu)^2 a_\nu'\Sigma_2^{-1}a_\nu\}.\end{aligned}$$

It is easy to see that $E_0(\mathrm{tr}\Sigma_2^{-1}\, a_\nu a_\nu') = E_0|\Sigma\sigma^{ij}(\nu_{1,j} - 1)^2)] = \mathrm{tr}\Sigma_2^{-1}/2$. Hence the conditional Fisher information is

$$\begin{aligned}E_0(-\ell_{\alpha\alpha}|Y_1) = \sum_n \{&E_0\mathrm{tr}[\Sigma_2^{-1}\dot\Sigma_{2,\nu}\Sigma_2^{-1}\dot\Sigma_{2,\nu}]\\ &+ (Y_1 - \mu)^2\mathrm{tr}\Sigma_2^{-1}\}/2,\end{aligned} \qquad [14]$$

and some additional calculation along the lines of Tang and Siegmund (11) shows that $\mathrm{tr}\Sigma_2^{-1} = (s-1)[1 + (s-2)\rho]/\{(1-\rho)[1 + (s-1)\rho]\}$.

The asymptotic conditional noncentrality parameter is

$$\alpha_0[E_0(-\ell_{\alpha\alpha}|Y_1)]^{1/2}. \qquad [15]$$

The expectation on the right-hand side of Eq. **14** can be evaluated by direct calculations or indirectly by observing that for random ascertainment the expected value of Eq. **14** must equal the unconditional Fisher information, which is simply the factor multiplying α_0^2/σ_Y^2 in Eq. **10**. (Recall that we are now taking $\sigma_Y^2 = 1$ for notational convenience.) In particular,

$$\begin{aligned}E_0\mathrm{tr}[\Sigma_2^{-1}\dot\Sigma_{2,\nu}\Sigma_2^{-1}\dot\Sigma_{2,\nu}] = &\binom{s}{2}\frac{\{[1 + (s-2)\rho]^2 + \rho^2\}}{\{(1-\rho)[1 + (s-1)\rho]\}^2}\\ &- \frac{(s-1)[1 + (s-2)\rho]}{\{(1-\rho)[1 + (s-1)\rho]\}}.\end{aligned} \qquad [16]$$

Because this expression is somewhat complicated for general s, we specialize to $s = 2$, for which we obtain

$$E_0(-\ell_{\alpha\alpha}|Y_1) = \sum 1\{Y_1 \in S\}[\rho^2/(1-\rho^2)^2 + (Y_1 - \mu)^2/2(1-\rho^2)], \qquad [17]$$

where S denotes the set of phenotypes for which a proband is ascertained. In large samples, by the law of large numbers the frequency of ascertained sibships converges to $\mathrm{P}\{Y_1 \in S\}$, and the average value of $1\{Y_1 \in S\}(Y_1 - \mu)^2$ converges to $E[(Y_1 - \mu)^2; Y_1 \in S]$. Hence, the large sample noncentrality *per ascertained sibship* is

$$\alpha_0[\rho^2/(1-\rho^2)^2 + E[(Y_1 - \mu)^2|Y_1 \in S]/2(1-\rho^2)]^{1/2}, \qquad [18]$$

which is consistent with Eq. **10** when $s = 2$ and all sibships are ascertained.

For a simple numerical example, it follows from Eq. **18** that if ascertainment is based on the upper 10% of the population phenotype, the number of sib pairs that must be *genotyped* is roughly 1/3 as many in random sampling. For sibships of size $s = 4$, about 1/2 as many ascertained sibships must be genotyped as random sibships. This gain in genotyping efficiency is smaller with a less stringent ascertainment criterion and with larger sibships.

Observe that, although we have begun the preceding analysis from an analytic expression for the conditional log likelihood given Y_1, one could equally well begin by writing the conditional log likelihood in the form of the sum of the unconditional log likelihood given in Eq. **5** and the negative log of the marginal probability density function of Y_1. Since this marginal probability does not depend on the genetic parameters α_0, δ_0, the efficient score ℓ_α has the same form as in the case of random ascertainment (i.e., the expressions in Eq. **6** evaluated at $\alpha_0 = 0$ and in Eq. **13** are equal), but the estimates of segregation parameters that enter into the final statistic are now determined by conditioning on Y_1.

In the case that we condition on the event that a sibship is ascertained, i.e., that $Y_1 \in S$, rather than the value of Y_1, the analysis is almost the same. The log likelihood function will now equal the sum of Eq. **5** and the additional term $-\log(\mathrm{P}\{Y_1 \in S\})$; but since the distribution of Y_1 involves only the segregation parameters, μ, ρ, and σ_Y^2, the efficient score ℓ_α is again unchanged. The efficient scores for the segregation parameters will change, but they are still uncorrelated with ℓ_α when $\alpha_0 = 0$. Consequently, the estimates of the segregation parameters will be different, but the asymptotic noncentrality parameter is still given by Eq. **16**. Note, however, that when conditioning on the exact phenotypic values of the ascertained siblings, the number r of ascertained siblings must be less than s, whereas in principle an ascertainment rule can involve all siblings if one conditions on the event of ascertainment. Thus, Risch and Zhang (28) discuss an ascertainment rule that involves both siblings of a sib pair, but their method has the disadvantage that it is most efficient when ascertainment involves fairly extreme phenotypes, and it does not extend in an obvious way to larger sibships.

Discussion of Ascertainment Corrections

In this article we have described a components of variance method of linkage analysis in sibships when sibships are either randomly ascertained or ascertained through a single proband. The method in principle is easily adapted to pedigrees other than sibships, although explicit results can be obtained in only a few special cases.

The most serious impediment to use of ascertainment corrections is lack of knowledge of the true ascertainment rule. To

some extent this problem is mitigated by conditioning on the exact phenotypic value of the proband(s), but this just removes the problem to the definition of the proband(s). For example, if the proband is identified through diagnosis of a disease related to the quantitative phenotype, should the ascertainment event be (as we have assumed) that a particular sib has the disease, that at least one sib has the disease, or something in between?

An appealing design that avoids some of these fundamental conceptual difficulties is to ascertain nuclear families through parents. This design may often involve both parents, but if a trait is of primary interest in only one sex, e.g., bone mineral density as a quantitative trait in women as it relates to osteoporosis, ascertainment through a particular parent may be relevant. In such a case, the analysis is simpler than that given above, since the conditional means and covariances do not depend on the number of alleles inherited identical by descent between proband and offspring, which is always one. The noncentrality parameter is of the same form as Eq. **10**, but with the sib correction ρ replaced by the conditional correlation $\rho - \tilde{\rho}^2$, where $\tilde{\rho}$ is the phenotypic correlation of parent and offspring. For traits that are purely additive and have no shared environmental covariance, $\rho = \tilde{\rho}$.

The normality assumption, which yields simple formulas for the conditional phenotypic expectations, variances and covariances given the phenotypes of ascertained relatives, plays an important role in the ascertainment corrections obtained in this article. The robustness of the resulting procedures and how they might be modified to become more robust to violations of the normality assumption and, perhaps even more importantly, to violations of the assumed method of ascertainment should be studied and reported in detail.

This research was supported by National Institutes of Health Grant R01 HG00849-09 and by a Stanford Graduate Fellowship.

1. Haseman, J. K. & Elston, R. C. (1972) *Behav. Genet.* **2,** 3–19.
2. Hopper, J. L. & Mathews, J. D. (1982) *Ann. Hum. Genet.* **46,** 373–383.
3. Amos, C. I. (1994) *Am. J. Hum. Genet.* **54,** 535–543.
4. Fulker, D. W. & Cherny, S. S. (1996) *Behav. Genet.* **26,** 527–532.
5. Almasy, L. & Blangero, J. (1998) *Am. J. Hum. Genet.* **62,** 1198–1211.
6. Williams, J. T. & Blangero J. (1999) *Ann. Hum. Genet.* **63,** 545–563.
7. Teng, J. (1996) Ph.D. thesis (Stanford University, Stanford, CA).
8. Wright, F. (1997) *Am. J. Hum. Genet.* **60,** 740–742.
9. Sham, P. C., Purcell, S., Cherny, S. S. & Abecasis, G. R. (2002) *Am. J. Hum. Genet.* **71,** 238–253.
10. Feingold, E. (2002) *Am. J. Hum. Genet.* **71,** 217–222.
11. Tang, H.-K. & Siegmund, D. (2001) *Biostatistics* **2,** 147–162.
12. Elston, R. C. & Sobel, E. (1979) *Am. J. Hum. Genet.* **31,** 62–69.
13. Fisher, R. A. (1918) *Trans. R. Soc. Edinburgh* **52,** 399–433.
14. Kempthorne, O. (1957) *Genetic Statistics* (Wiley, New York).
15. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. (1996) *Am. J. Hum. Genet.* **58,** 1347–1363.
16. Teng, J. & Siegmund, D. (1999) *Biometrics* **54,** 379–411.
17. Feingold, E., Brown, P. O. & Siegmund, D. (1993) *Am. J. Hum. Genet.* **53,** 234–251.
18. Siegmund, D. (1985) *Sequential Analysis: Tests and Confidence Intervals* (Springer, New York).
19. Lander, E. S. & Kruglyak, L. (1995) *Nat. Genet.* **11,** 241–247.
20. Elston, R., Buxbaum, S., Jacobs, K. B. & Olson, J. M. (2000) *Genet. Epidemiol.* **19,** 1–17.
21. Xu, X., Weiss, S., Xu, X. & Wei, L. J. (2000) *Am. J. Hum. Genet.* **67,** 1025–1028.
22. T.Cuenco, K., Szatkiewicz, J. P. & Feingold, E. (2003) *Am. J. Hum. Genet.* **73,** 863–873.
23. Lange, K. T., Little, R. J. A. & Taylor, J. M. G. (1989) *J. Am. Stat. Assoc.* **84,** 881–896.
24. Dupuis, J. & Siegmund, D. (2000) *Game Theory, Optimal Stopping, Probability and Statistics*, eds. F. Thomas Bruss and L. Le Cam (Institute of Mathematical Statistics, Hayward, CA) pp. 141–152.
25. Wang, K. (2003) *Hum. Hered.* **55,** 1–15.
26. Beaty, T. H. & Liang, K. Y. (1987) *Genet. Epidemiol.* **4,** 203–210.
27. Andrade, M. & Amos, C. I. (2000) *Genet. Epidemiol.* **19,** 333–344.
28. Risch, N. & Zhang, H. (1995) *Science* **268,** 1584–1589.