Am. J. Hum. Genet. 49:786–796, 1991

Linkage Analysis of Quantitative Traits: Increased Power by Using Selected Samples

Gregory Carey* and John Williamson†

*Institute for Behavioral Genetics and Department of Psychology, and †Department of Mathematics, University of Colorado, Boulder

Summary

Although a number of methods have been developed for linkage analysis of quantitative traits, power is relatively poor unless there is a single major locus of very large effect. Here it is demonstrated that the use of selected samples (i.e., ascertainment of a proband with an extreme score on the quantitative measure) can dramatically increase power, especially when proband selection is performed on the tail of a distribution with an infrequent recessive gene. Depending on gene action and allele frequency, selected samples permit detection of a major locus that accounts for as little as 10%–20% of the phenotypic variation. The judicious use of selected samples can make an appreciable difference in the feasibility of linkage studies for quantitative traits.

Introduction

Many traits of importance to human and medical genetics are quantitative. Plasma glucose and cholesterol are two clear medical examples, but virtually any measure of enzyme activity or receptor binding is indexed on a continuous scale. Linkage methods have been developed for continuous traits, most of which are samples from the general population (Haseman and Elston 1972; Smith 1975; Lange et al. 1976; Blackwelder and Elston 1982; Cockerham and Weir 1983; Amos and Elston 1989; Amos et al. 1989; Nance and Neale 1989; Goldgar 1990). Power calculations suggest that linkage may be detected when there is a locus of very large effect, e.g., one responsible for 50% or more of the variation. With a more modest major-locus effect, the required number of pairs in a sib-pair linkage analysis exceeds 1,000, making the establishment of cell lines and marker typing both expensive and time consuming.

Samples selected through a proband with an extreme score on a quantitative trait may sometimes increase statistical power relative to that for random samples (Jayakar et al. 1984; DeFries and Fulker 1988; Rao et al. 1988; Boehnke and Moll 1989; Demenais and Amos 1989; R. C. Elston, personal communication). In its extreme form, a selected sample for a quantitative trait would ascertain affected relative pairs by, say, requiring that each member of a sib pair have a score exceeding a certain threshold value. In its less extreme form, only one relative would be ascertained. Hence, it is worthwhile exploring the extent to which selection might increase the power to detect linkage. Here we demonstrate that selection can *dramatically* increase power, in some cases by an order of magnitude. Moreover, the selected-sample method can have sufficient power to detect major-locus effects responsible for as little as 10%–20% of phenotypic variance. We illustrate these principles by using the sib-pair method.

Received April 8, 1991; revision received June 10, 1991.

Address for correspondence and reprints: Gregory Carey, Institute for Behavioral Genetics, University of Colorado, Box 447, Boulder, CO 80309-0447.

0002-9297/91/4904-0011$02.00

Material and Methods

A Model of Linkage in Sib Pairs

To examine the utility of selected samples, models are required for sibling similarity both in random samples of sibs and in selected samples. Furthermore, the method used in selected samples must be analogous to that used in unselected samples. Otherwise, differences in random versus selected samples may be due

to the method of analysis rather than to actual power differences between the two sampling strategies. While there are appropriate methods developed for random samples and for selected samples, we know of no methods that are analogous. Here, we develop analogous methods. Specifically, we propose the use of linear models to test for differences in the slope and intercept of the regression lines between one sib's score and a second sib's score, as a function of identity by descent (ibd) at a marker. We acknowledge that, when there is a major locus of large effect, the regression of one sib's score on the other sib's score is nonlinear. Nevertheless, we demonstrate in the Appendix that the coefficients of a linear model provide a valid and robust test for linkage. Because the focus here is on selection, we simplify exposition by assuming that marker ibd can be unequivocally determined.

Let the two members of a sib pair be termed the X and Y sib, with respective phenotypic scores of x and y and correlation coefficient ρ. We assume that x and y, conditional on the major genotypes of sib X and sib Y, are distributed as a bivariate normal with mean vector $(\mu + g_x, \mu + g_y)^t$, where g_x and g_y are deviations for the major genotypes of sib X and Y, respectively. The covariance matrix may be written as

$$\sigma^2_w \begin{pmatrix} 1 & \omega \\ \omega & 1 \end{pmatrix}.$$

Let δ denote a variable with values of $-1/2$, 0, and $1/2$ for sib pairs who share zero, one, and two alleles ibd at a marker, respectively. Let γ denote a second variable with values of $1/4$, $-1/4$, and $1/4$ for pairs with, respectively, zero, one, and two alleles ibd at the marker. Let δx be the product of sib X's phenotypic score and the pair's delta value, and let γx be the product of x and the pair's γ value. Let b_i denote a regression coefficient, and let u denote a residual. Then the linear equation we propose is $y = b_1 + b_2\delta + b_3\gamma + b_4x + b_5\delta x + b_6\gamma x + u$. In matrix notation, let $\mathbf{y}$ denote the vector of y scores, X_n denote the design matrix, $\mathbf{b}$ denote the vector of regression coefficients, and $\mathbf{u}$ denote the vector of residuals. The model may now be written as $\mathbf{y} = X_n\mathbf{b} + \mathbf{u}$, so the least-squares solution for $\mathbf{b}$ will be $\hat{\mathbf{b}} = (X_n^tX_n)^{-1}X_n^t\mathbf{y}$. As the sample size grows large, the random entries in the 6×6 matrix $(1/n)X_n^tX_n$ converge to constant values. For any fixed sample size, n, these elements are equal to the expectations of the corresponding elements in $(1/n)X_n^tX_n$, or

$$\begin{pmatrix} 1 & 0 & 0 & \mu & 0 & 0 \\ 0 & 1/8 & 0 & 0 & \mu/8 & 0 \\ 0 & 0 & 1/16 & 0 & 0 & \mu/16 \\ \mu & 0 & 0 & E(x^2) & 0 & 0 \\ 0 & \mu/8 & 0 & 0 & E(x^2)/8 & 0 \\ 0 & 0 & \mu/16 & 0 & 0 & E(x^2)/16 \end{pmatrix},$$

which we denote by M. Similarly, the entries in the vector $\mathbf{c} = E[(1/n)X_n^t\mathbf{y}]$ are

$$\{\mu,0,0,E(xy),[E(xy|ibd_m = 2) - E(xy|ibd_m = 0)]/8,[E(xy|ibd_m = 0) + E(xy|ibd_m = 2) - 2E(xy|ibd_m = 1)]/16\}^t.$$

Tedious algebra may be used to invert M, postmultiply by $\mathbf{c}$, and obtain the values of $\hat{\mathbf{b}}$. They are presented in the second column of table 1.

Table 1

Asymptotic Expectations of Coefficients for Linear Model of Sib-Pair Linkage for Quantitative Trait

Quantity	Form of $\mathbf{b}$ from Solving $\hat{\mathbf{b}} = M^{-1}\mathbf{c}$	$\mathbf{b}$ Expressed in Terms of Genetic Parameters
b_1	$\mu(1-\rho)$	$\mu(1-\rho)$
b_2	$-\mu[cov(x,y\|ibd_m = 2) - cov(x,y\|ibd_m = 0)]/\sigma^2$	$-\mu rH$
b_3	$-\mu[cov(x,y\|ibd_m = 0) + cov(x,y\|ibd_m = 2) - 2cov(x,y\|ibd_m = 1)]/\sigma^2$	$-\mu r^2D$
b_4	ρ	ρ
b_5	$[cov(x,y\|ibd_m = 2) - cov(x,y\|ibd_m = 0)]/\sigma^2$	rH
b_6	$[cov(x,y\|ibd_m = 0) + cov(x,y\|ibd_m = 2) - 2cov(x,y\|ibd_m = 1)]/\sigma^2$	r^2d

NOTE.—μ = population mean; ρ = correlation coefficient; σ = phenotypic SD; ibd_m = number of alleles identical by descent at a marker; r = correlation in ibd bedtween marker and quantitative locus; H = broad-sense heritability; D = proportion of phenotypic variance due to dominance variance.

We may now apply genetic theory to express the elements of **b** in terms of genetic parameters. Write the phenotype as a linear combination of the population mean (μ), a deviation due to genotype at the major locus (g), and a deviation within major-locus genotype (w). Thus, $x = \mu + g_x + w_x$ and $y = \mu + g_y + w_y$. We assume that mating is random with respect to the trait and that deviation scores within major genotypes are uncorrelated with genotypic values. Then, phenotypic variance and sibling covariance become $\sigma^2 = \sigma_a^2 + \sigma_d^2 + \sigma_w^2$ and $\text{cov}(x,y) = \frac{1}{2}\sigma_a^2 + \frac{1}{4}\sigma_d^2 + \omega\sigma_w^2 = \rho\sigma^2$, where σ_a^2 is additive genetic variance at the major locus, σ_d^2 is dominance variance, σ_w^2 is variance within the major locus, ω is the intraclass correlation for "background" factors (i.e., deviations within a major genotype), and ρ is the sibling intraclass correlation coefficient for the phenotypic scores. Note that ω will be a function of both background genetic factors (i.e., loci other than the quantitative-trait locus [QTL]) and environmental factors.

Conditioning the sibling covariance on ibd at the QTL (ibd_q) gives $\text{cov}(x,y|\text{ibd}_q = 0) = \omega\sigma_w^2$, $\text{cov}(x,y|\text{ibd}_q = 1) = \omega\sigma_w^2 + \frac{1}{2}\sigma_a^2$, and $\text{cov}(x,y|\text{ibd}_q = 2) = \omega\sigma_w^2 + \sigma_a^2 + \sigma_d^2$. The covariances conditional on ibd at a marker (ibd_m) are

$$\text{cov}(x,y|\text{ibd}_m = i) = \sum_{j=0}^{2} \text{prob}(\text{ibd}_q = j|\text{ibd}_m = i)\,\text{cov}(x,y|\text{ibd}_q = j).$$

The quantity $\text{prob}(\text{ibd}_q = j|\text{ibd}_m = i)$, the conditional probability that a sib pair share j alleles ibd at the quantitative locus, given that they share i alleles ibd at the marker, has been previously tabulated (Haseman and Elston 1972; Bishop and Williamson 1990) under equal recombination in males and females, an assumption we also make here. Let θ denote the recombination fraction between the major locus and marker locus, and let $T = \theta^2 + (1-\theta)^2$. The conditional sib covariance, given 0 alleles ibd at the marker, becomes

$$\begin{aligned}\text{cov}(x,y|\text{ibd}_m = 0) &= T^2\omega\sigma_w^2 + 2T(1-T)(\omega\sigma_w^2 + \tfrac{1}{2}\sigma_a^2) + (1-T)^2(\omega\sigma_w^2 + \sigma_a^2 + \sigma_d^2) = \\ &\omega\sigma_w^2 + (1-T)\sigma_a^2 + (1-T)^2\sigma_d^2 = \omega\sigma_w^2 + 2\theta(1-\theta)\sigma_a^2 + 4\theta^2(1-\theta)^2\sigma_d^2 = \omega\sigma_w^2 + \tfrac{1}{2}\sigma_a^2 + \tfrac{1}{4}\sigma_d^2 - \\ &\tfrac{1}{2}(1-2\theta)^2(\sigma_a^2 + \sigma_d^2) + \tfrac{1}{4}(1-2\theta)^4\sigma_d^2 = \rho\sigma^2 - \tfrac{1}{2}r\sigma_g^2 + \tfrac{1}{4}r^2\sigma_d^2,\end{aligned}$$

where $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$ and $r = (1-2\theta)^2 = \text{corr}(\text{ibd}_m, \text{ibd}_q)$, or the correlation between ibd at the marker and ibd at the QTL. When similar algebra is used, the other two covariances conditional on marker ibd may be written as $\text{cov}(x,y|\text{ibd}_m = 1) = (\rho\sigma^2 - \frac{1}{4}r^2\sigma_d^2)$ and $\text{cov}(x,y|\text{ibd}_m = 2) = \rho\sigma^2 + \frac{1}{2}r\sigma_g^2 + \frac{1}{4}r^2\sigma_d^2)$. Substitution of these three covariances into the second column of table 1 gives the sib b's in terms of the genetic parameters. They are listed in the third column of table 1.

As a test for linkage, we propose to compare the least-squares estimate of b_5 against its standard error (SE). The quantity b_5 is a product of the proportion of phenotypic variance due to the major locus (H, or the broad-sense heritability) and to the correlational distance between the major locus and the marker (r), under the assumption of Hardy-Weinberg equilibrium and linkage equilibrium. If there is no linkage, then $r = 0$ and $b_5 = 0$.

We now develop a similar linear model for selected samples. It is assumed that one and only one member of a sib pair, designated here as the X sib, is ascertained as a proband because of a high trait score and that only one sibling, denoted as the Y sib, is studied. In this case, selection is independent of marker type and ibd values. Hence, the distribution of major genotypes is the same among $\text{ibd}_q = 0$ probands as it is among $\text{ibd}_q = 1$ and $\text{ibd}_q = 2$ probands.

Let an asterisk (*) denote a variable or a function conditional on selection on x. Hence, if $\phi(x)$ is the density function for the X sib in the unselected population, then $\phi(x^*)$ will denote the density function in the selected population. The linear regression model given in the unselected population may now be modified to detect linkage in a selected population: $y^* = b_1^* + b_2^*x^* + b_3^*\delta + b_4^*\gamma + u^*$. Now we test for a difference in means as a function of ibd instead of for a difference in slope.

Let X_n^* denote the $n \times 4$ matrix with the ith row being $(1, x_i, \delta_i, \gamma_i)$ for the ith sib pair. Let M^* denote the 4×4 matrix $E([1/n]X_n^{*t}X_n^*)$. In large samples, the elements of $([1/n]X_n^{*t}X_n^*)$ can be approximated by $M^* =$

$$\begin{pmatrix} 1 & \mu_{x^*} & 0 & 0 \\ \mu_{x^*} & \sigma_{x^*}^2 + \mu_{x^*}^2 & 0 & 0 \\ 0 & 0 & 1/8 & 0 \\ 0 & 0 & 0 & 1/16 \end{pmatrix},$$

where μ_{x^*} is the mean quantitative score for the selected probands and $\sigma^2_{x^*}$ is the variance. Let $\mathbf{c}^* = E[(1/n)X_n^{*t}y^*] =$

$$\{\mu_{y^*}, E(x^*y^*), [E(y^*|\text{ibd}_m = 2) - E(y^*|\text{ibd}_m = 0)]/8, [E(y^*|\text{ibd}_m = 1) + E(y^*|\text{ibd}_m = 2) - 2E(y^*|\text{ibd}_m = 0)]/16\}^t.$$

As sample size grows large $(1/n)X^{*t}_n y^*$ converges to the constant vector $\mathbf{c}^*$ so that in large samples $(1/n)X^{*t}_n y^*$ can be approximated by $\mathbf{c}^*$. The values of vector $\mathbf{b}^* = M^{*-1}\mathbf{c}^*$ are given in table 2.

We direct attention to the coefficient b^*_3 in table 2. This equals the expectation of the Y sib, given two alleles ibd at the marker, less the expectation of the Y sib, given zero alleles at the marker. Before expressing this expectation in terms of genetic parameters, we digress for a moment to develop notation.

Selection on x will change the frequency of the ith genotype, from f_i in the general population to f^*_i in the selected population, $i = (1, 2, 3)$, corresponding to genotypes aa, Aa, and AA. Selection will also change the background mean value for probands with the ith major genotype, from $\overline{w}_i$ which equals 0.0 to $\overline{w}^*_i$. Hence, the proband mean may be written as

$$\mu_{x^*} = \mu + \sum_i f^*_i(g_i + \overline{w}^*_i) = \mu + g^*_x + \overline{w}_{x^*}.$$

Let E_k denote the mean value for siblings with k alleles ibd at the QTL, and let S_{jik} denote the probability that a sibling will have the jth major genotype, given that the proband has the ith major genotype and given that the pair share k alleles at the QTL. Then

$$E_k = \sum_i f^*_i \sum_j S_{jik}(\mu + g_j + \omega\overline{w}^*_i).$$

Let p be the frequency of the decreaser allele, with $q = 1 - p$ being the frequency of the increaser allele. Let subscripts 1, 2, and 3 denote genotypes aa, Aa, and AA, respectively, and recall that ω denotes the intraclass correlation for background factors. Tedious algebra then gives the expectations for siblings, as a function of ibd at the QTL, as

$$E_0 = \mu + \omega\overline{w}_{x^*},$$

$$E_1 = \mu + p(f^*_1 + \tfrac{1}{2}f^*_2)g_1 + (qf^*_1 + \tfrac{1}{2} + pf^*_3)g_2 + q(\tfrac{1}{2}f^*_2 + f^*_3)g_3 + \omega\overline{w}_{x^*},$$

and

$$E_2 = \mu + f^*_1 g_1 + f^*_2 g_2 + f^*_3 g_3 + \omega\overline{w}_{x^*}.$$

Then the expectation of y^*, conditional on marker ibd, is

$$E(y^*|\text{ibd}_m = i) = \sum_{j=0}^{2} \text{prob}\{\text{ibd}_q = j|\text{ibd}_m = i\}E_j,$$

giving

$$E(y^*|\text{ibd}_m = 0) = T^2E_0 + 2T(1-T)E_1 + (1-T)^2E_2;$$

$$E(y^*|\text{ibd}_m = 1) = T(1-T)E_0 + [T^2 - (1-T)^2]E_1 + T(1-T)E_2;$$

$$E(y^*|\text{ibd}_m = 2) = (1-T)^2E_0 + 2T(1-T)E_1 + T^2E_2.$$

Hence, we may write coefficient b^*_3 as

$$b^*_3 = [(1-T)^2 - T^2](E_2 - E_0) = (1-2T)(E_2 - E_0) = r(f^*_1 g_1 + f^*_2 g_2 + f^*_3 g_3),$$

or the product of the correlational distance between the QTL and the marker and the average proband genotypic value at the QTL. If $r = 0$ (which, of course, implies no linkage) or if there is no major-locus effect, then $b^*_3 = 0$.

The above expectations are only asymptotically valid, and presentation of exact asymptotic SEs would take us into a complicated area beyond the main purpose of the present paper (see Appendix). Hence, we used Monte Carlo simulation to evaluate the power of selected versus unselected samples of small size. For the genetic model, we assume a diallelic major locus in Hardy-Weinberg equilibrium. We also assume that the trait scores of the X and Y siblings are bivariate normal conditional on their major genotypes. The model for selection assumed that probands had quantitative scores greater than a certain value, t, which was numerically estimated from the parameters of the genetic model and the percent selected. Two sample sizes—i.e., 240 pairs and 480 pairs—were used, and 100 replicates were generated for each set of parameter values. Regression was used to estimate the b's or the b^*'s and their SEs. The statistics $z = \hat{b}_5/\text{SE}(\hat{b}_5)$ in unselected samples and $z = \hat{b}^*_3/\text{SE}(\hat{b}^*_3)$ in the selected case were treated as normally distributed. Power at

Table 2

Asymptotic Expectations for Regression Coefficients in Selected Sample

$\mathbf{b}^*$	Form of $\mathbf{b}^*$ from Solving $\mathbf{b}^* = M^{*-1}\mathbf{c}^*$
b^*_1	$\mu_{y^*} - b_2\mu_{x^*}$
b^*_2	$\text{cov}(x^*, y^*)/\sigma^2_{x^*}$
b^*_3	$E(y\|\text{ibd}_m = 2) - E(y^*\|\text{ibd}_m = 0)$
b^*_4	$E(y^*\|\text{IBD}_m = 0) + E(y^*\|\text{ibd}_m = 2) - 2E(y^*\|\text{ibd}_m = 1)$

NOTE.—Symbols are as defined in table 1.

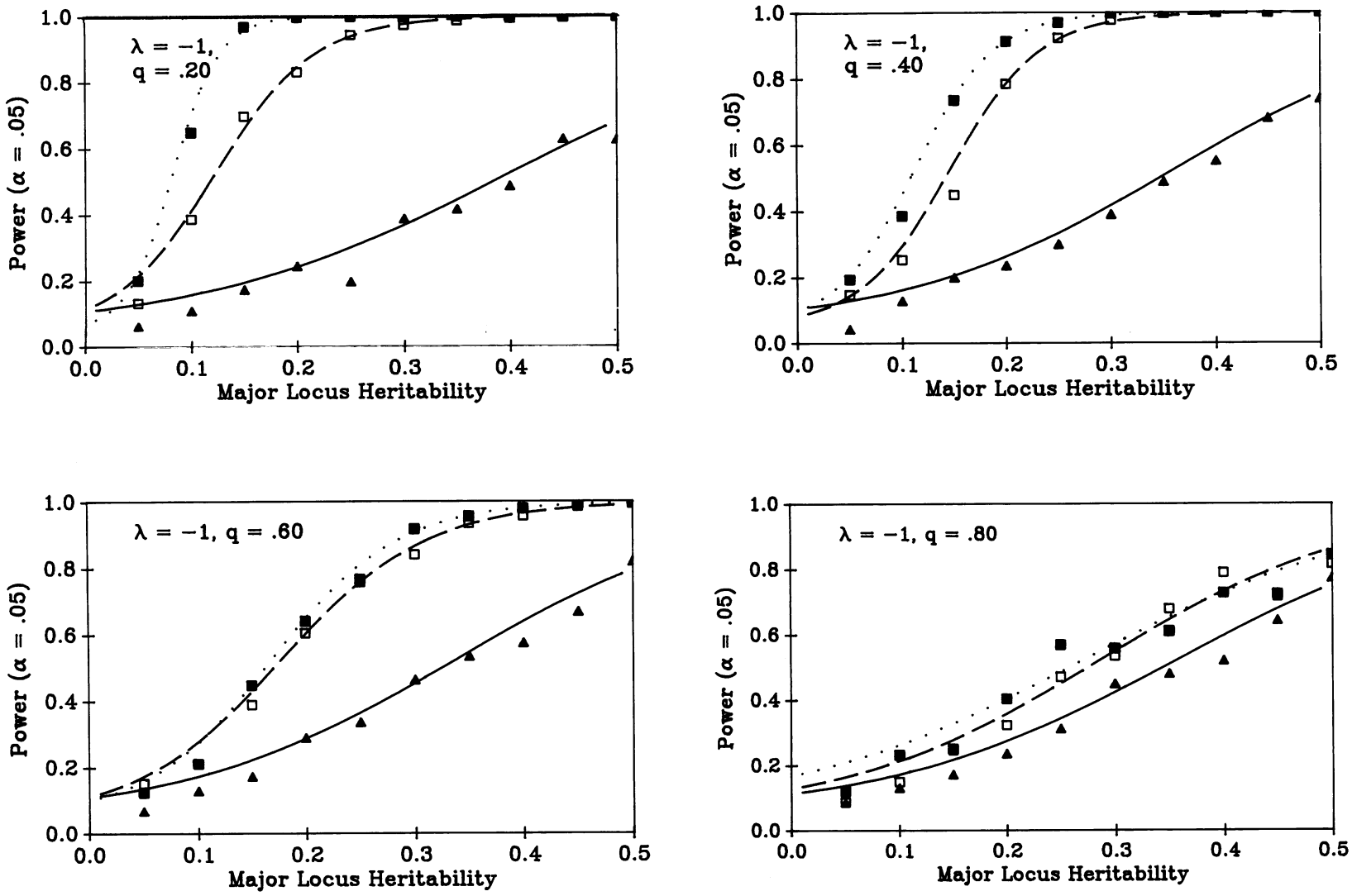

Figure 1 Power to detect linkage with 240 sibling pairs in random sample (▲ – ▲), selected sample when top 10% of distribution is used as probands (□- -□), and top 5% of distribution (■· · ·■): sampling tail with recessive allele with frequency q.

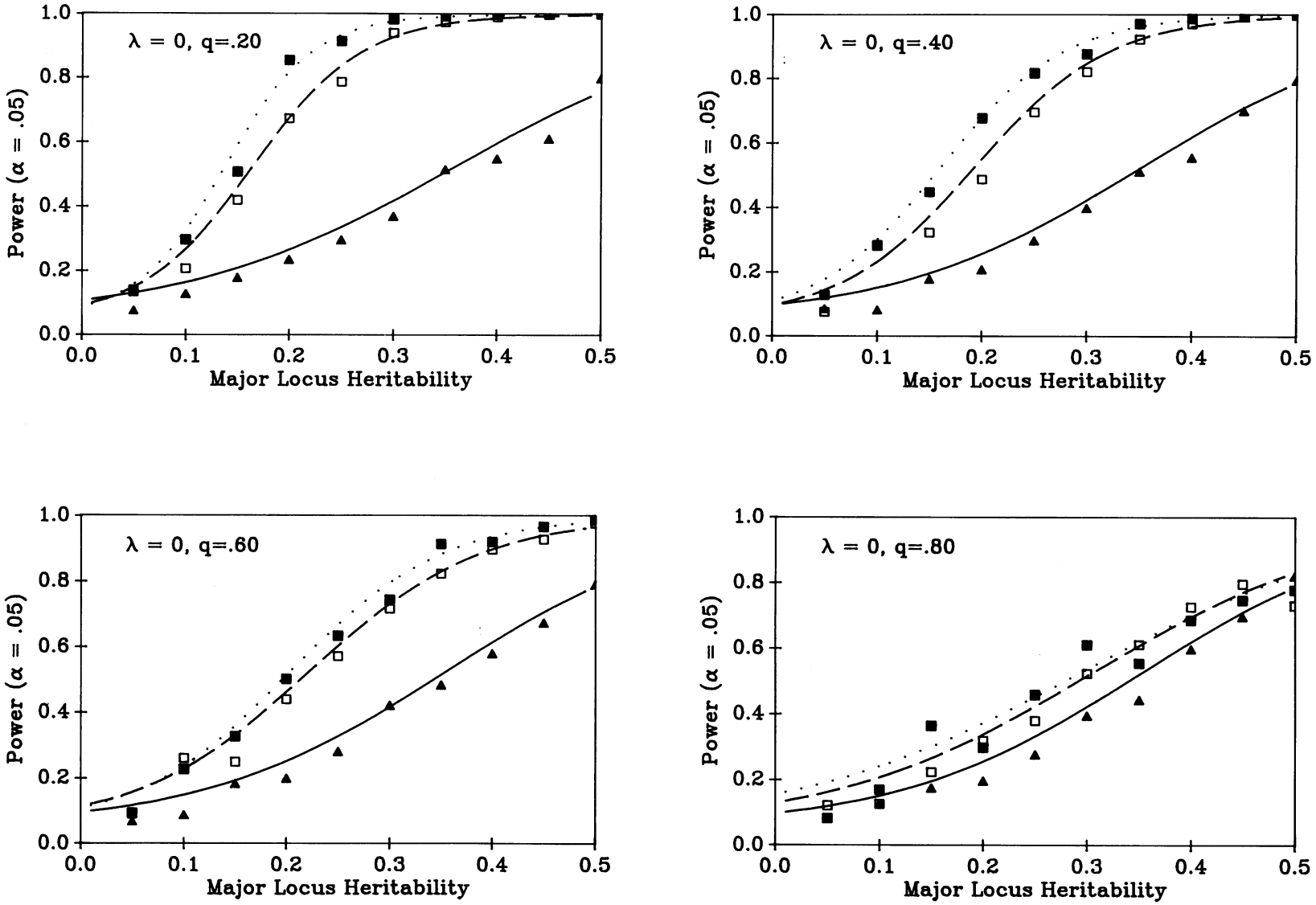

Figure 2 Power to detect linkage with 240 sibling pairs in random sample (▲ — ▲), selected sample when top 10% of distribution is used as probands (□- -□), and top 5% of distribution (■···■): additive gene action.

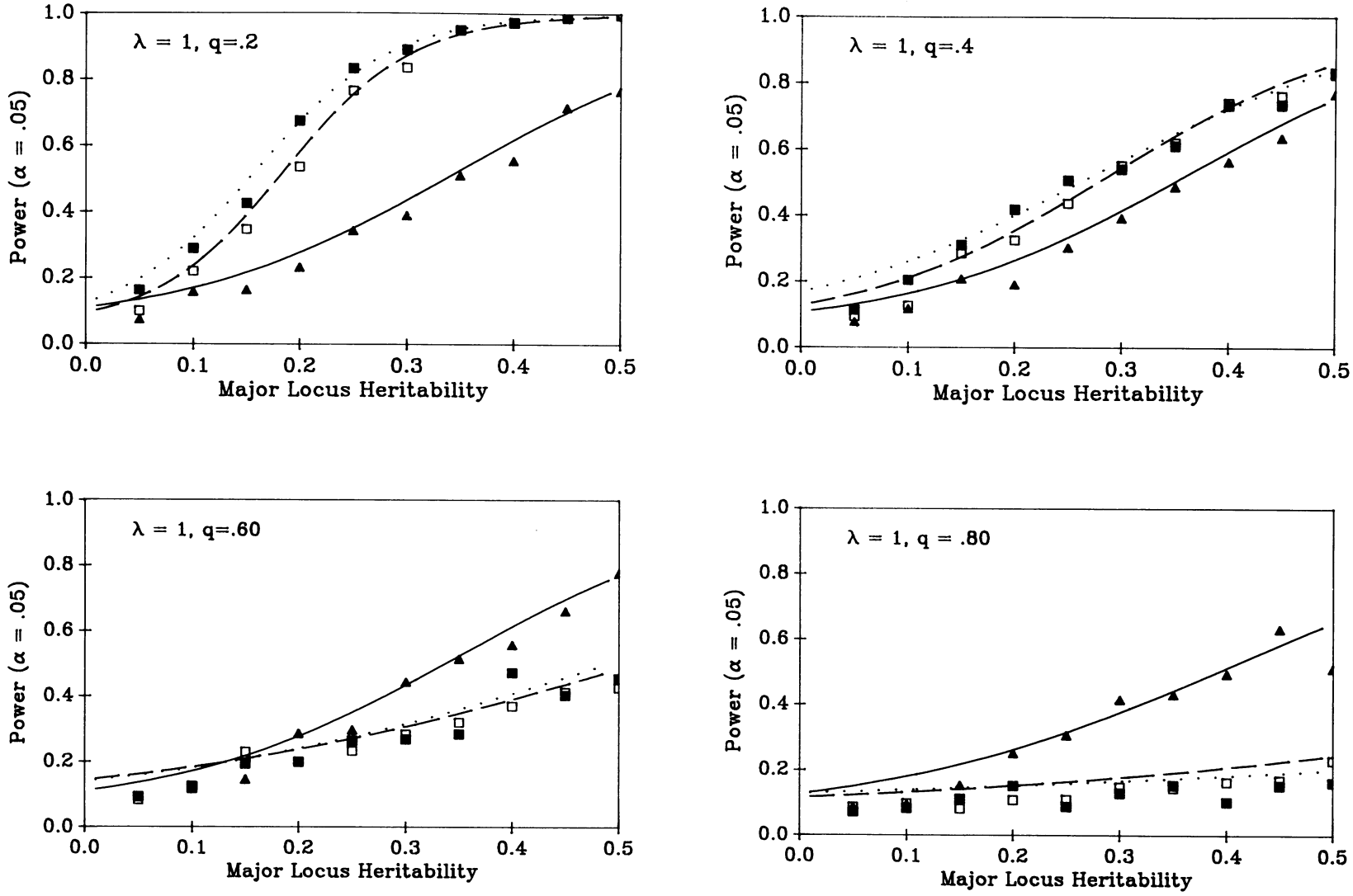

Figure 3 Power to detect linkage with 240 sibling pairs in random sample (▲ — ▲), selected sample when top 10% of distribution is used as probands (□- -□), and top 5% of distribution (■· · ·■): sampling tail with dominant allele with frequency q.

$\alpha = .05$ was assessed as the proportion of times z exceeded the one-tailed critical value of 1.645. Naturally, setting $\alpha = .05$ requires that a positive linkage result must be replicated in another laboratory.

In the unselected population, we set $\mu = 0$ and $\sigma_w^2 = 1$. The parameters were varied as follows: frequency of the decreaser allele = $p = .2, .4, .6$, and .8; $\theta = 0, .05$, and .10; $\omega = .2$ and .4; $H = .05$–.50 by .05 increments; and proportion selected = .05, .10, and 1.0. For each set of these parameter values, gene action was parameterized by $\lambda = 1$ (increaser allele A dominant to decreaser allele a), $\lambda = 0$ (additive gene action), and $\lambda = -1$ (decreaser allele a dominant to increaser allele A). Together, μ, p, H, λ, and σ_w^2 determine the values of g_i.

Results

Figures 1–3 present power curves for 240 sib pairs in which the increaser allele is recessive (fig. 1), additive (fig. 2), and dominant (fig. 3) under different allele frequencies for the increaser allele (q). Selection is done on the upper tail of the distribution, and the information is expressed in terms of major-locus heritability. The same information could be displayed in terms of deviations between the homozygotes. Smoothed lines are the logistic curves that best fit the Monte Carlo-ed data points. The figures were plotted for the situation in which ω (the sib correlation for background variance) is .20 and θ is .05 and allele frequency is expressed as $q = (1 - p)$ = frequency of the increaser allele.

When sampling is performed from the rare recessive tail of a distribution (fig. 1; $q = .2$), there is a dramatic increase in power when a selected sample strategy is used. The threshold for selection is also an important consideration in this case. Selecting the upper 5% as probands gives a reasonable chance for detecting linkage for a locus explaining as little as 10% of the phenotypic variance. As the QTL heritability increases, the difference between selecting the upper 5% versus the upper 10% becomes less important. In all cases, the selected samples increase power relative to random samples, although, as allele frequency increases, the difference in power curves diminishes.

Under additive gene action (fig. 2), selected samples also increase power under all allele frequencies, albeit not as dramatically as in the case of an infrequent recessive. Under additive gene action, there is only a small difference between the 5% and 10% selection differential.

With dominant gene action for the increaser allele, the power of selected sibling samples depends critically on both allele frequency and the tail of the distribution that is selected. Sampling from the tail of a distribution with a rare dominant increases power, but sampling from the tail with a common dominant actually decreases power.

Examination of figures 1–3 jointly gives one important corollary to the use of selected samples. Irrespective of the mode of gene action, the power differential between selected and random samples is greatest when one samples from the tail with the allele of lesser frequency. Clearly, if the mode of transmission and/or allele frequency are not known, then sampling sib probands from both ends of a continuous distribution must be considered. It is unclear how much this power diminution as a function of gene action and allele frequency holds for other linkage designs such as multigenerational pedigrees.

Table 3 gives the sample sizes required to detect a locus with 20% broad-sense heritability and 80% power. The sample sizes are based on the Monte Carlo results and hence are only approximations. In a random sample, it would take 1,400–1,600 sib pairs to achieve satisfactory power. In general, the selection strategy requires sample sizes in the hundreds, not the thousands—and, in some cases, fewer than 100 sib pairs would permit satisfactory tests for linkage. There are only two conditions in the selected sample that would require more pairs: (1) dominant gene action and (2) increaser allele frequency .60 or greater. Once again, this obstacle may be overcome by sampling from both tails of a distribution.

Discussion

Selected samples are an efficient method for detecting linkage with quantitative traits, because, depending on gene action, fewer sib pairs require marker typing than would be the case in a general population sample. When the major locus does not account for an overwhelming proportion of phenotypic variance, the increase in power from selection can make the difference between feasibility and infeasibility of a linkage study. The fact that, in the face of strong background variation, selected samples can detect major loci of moderate effect also opens avenues of investigation of linkage for quantitative traits that are correlated with disease liability.

The disadvantage, of course, is finding extreme probands in the first place. In many medical applications,

Table 3

Approximate Number of Sib Pairs, in Selected Samples, Required to Detect Linkage between Marker of Known ibd and Major Quantitative Locus with Recombination Fraction .05

	Approximate No. of Sib Pairs in					
	Upper-5% Selected Sample			Upper-10% Selected Sample		
q	R	A	D	R	A	D
.20	73	261	350	199	328	467
.40	180	358	786	251	559	889
.60	358	643	1,710	445	708	1,843
.80	762	1,010	7,009	1,044	1,032	8,582

Note.—Power is set at 80% with $\alpha = .05$. The example is for a major locus contributing to 20% of phenotypic variation. In a random sample, approximately 1,500 pairs would be required. q = frequency of the increaser allele; R = recessive gene action for increaser allele; A = additive gene action for increaser allele; D = dominant gene action for increaser allele. Other symbols are as defined in table 1.

probands with extreme scores on fasting plasma glucose, plasma cholesterol, hypertension, etc. may be easily ascertained through clinics. Indeed, with traits correlated with fitness, one would expect to find rare recessives at the deleterious tail of the distribution. In other circumstances, the utility of the design must balance the cost of measuring phenotypes against the expense of establishing cell lines and/or typing markers.

In either event, the fact that gene action and allele frequency moderate the power of selected sibling samples dictates that other techniques of genetic epidemiology, particularly segregation and commingling analyses, must play a strong preparatory role in the design of a selected-sample linkage study. The reason why gene action moderates power is unclear. We suspect that one major statistic strongly correlated with power is the displacement of the two homozygotes. For a trait with a given heritability, this displacement will be greater for a genetic system in which there is a rare recessive locus than it is for one in which there is a rare dominant.

Because the main purpose of the present paper has been to demonstrate the utility of selected samples, the methods have been developed to give expectations that are easily solved using numerical methods. While these methods are robust in the sense that the parameters of a major-locus model are not required to detect linkage, they may be less powerful than other analytical techniques. Also unexplored are other avenues of the selected-sample strategy, such as (*a*) varying the selection threshold as a function of segregation parameters, to optimize linkage strategies; (*b*) sampling from both ends of a distribution; and (*c*) sampling "concordant" pairs. It is clear that the method explored by Boehnke and Moll (1989)—i.e., sequential sampling of families to isolate those for whom there is evidence for major-locus segregation—should be utilized in future research designs.

Acknowledgments

This work was supported in part by NIDA grant DA05131. We thank David W. Fulker and John C. DeFries for the suggestion of using regression to study sib-pair linkage and for comments on an earlier draft of the present paper. We also thank two anonymous referees for their helpful comments.

Appendix

For the unselected model consider the equation $y = b_1 + b_2\delta + b_3\gamma + b_4x + b_5\delta x + b_6\gamma x + u$. Form the 6×6 matrix, M, in the following way. The first row of M is the row obtained by taking expectations of the components in the row vector $(1,\delta,\gamma,x,\delta x,\gamma x)$. Next multiply each component of this row vector by δ and again take expectations to obtain the second row of M. The third, fourth, fifth, and sixth rows of M are obtained in the same manner: the initial row, $(1,\delta,\gamma,x,\delta x,\gamma x)$, is multiplied in turn by γ, x, δx, and γx, and, in each row, expectations are taken. Form the column vector $\mathbf{c}$ by setting $\mathbf{c} = \text{col}[E(y),E(\delta y), E(\gamma y),E(\delta xy),E(\gamma xy)]$, and let $\mathbf{b} = \text{col}(b_1, b_2, b_3, b_4,$

b_5, b_6). If $\mathbf{b}$ satisfies $\mathbf{c} = M\mathbf{b}$, then u will be uncorrelated with each of the random variables δ, γ, x, δx, γx, and $E(u) = 0$.

Next let X_n be the $n \times 6$ matrix whose jth row is the vector of observations $(1, \delta_j, \gamma_j, x_j, \gamma_j x_j)$ from the jth sibling pair in the random sample of size n. If we let $\hat{\mathbf{b}}$ denote the least-squares estimator of $\mathbf{b}$, then $\hat{\mathbf{b}} = (X_n^t X_n)^{-1} X_n^t \mathbf{y}$, where $\mathbf{y} = \text{col}(y_1, y_2, \ldots, y_n)$ is the vector of phenotypic values for the Y siblings in the sample.

As mentioned in the main text, we are not working with a linear model, so the standard results concerning the unbiasedness and consistency of $\hat{\mathbf{b}}$ cannot be invoked. However, $\lim_{n\to\infty} n(X_n^t X_n)^{-1} = M^{-1}$ with probability 1 and $\lim_{n\to\infty} (1/n) X_n^t \mathbf{y} = \mathbf{c}$ with probability 1, so that we have the following result:

Result 1

In the absence of linearity it is still true that $\lim_{n\to\infty} \hat{\mathbf{b}} = M^{-1}\mathbf{c} = \mathbf{b}$ with probability 1. We can restate this result by saying that, as the number of sibling pairs being sampled grows large, $\hat{\mathbf{b}}$ will converge to $\mathbf{b}$. That is, $\hat{\mathbf{b}}$ is a consistent estimator of $\mathbf{b}$.

To obtain asymptotic formulas for the SEs of these least-squares estimators, write $\lim_{n\to\infty} E[\sqrt{n}(\mathbf{b}-\hat{\mathbf{b}})\sqrt{n}(\mathbf{b}-\hat{\mathbf{b}})^t] = \lim_{n\to\infty} E[n(X_n^t X_n)^{-1} X_n^t (\mathbf{y} - X_n\mathbf{b})(\mathbf{y} - X_n\mathbf{b})^t X_n (X_n^t X_n)^{-1} = \lim_{n\to\infty} E[n(X_n^t X_n)^{-1} X_n^t \mathbf{u}\mathbf{u}^t X_n (X_n^t X_n)^{-1}]$, where $\mathbf{u}$ is the vector of residuals for the n sibling-pair observations. In the standard linear model the assumption of homoscedasticity permits us to first condition on the values in the matrix X_n and then move the expectation through to the matrix $\mathbf{u}\mathbf{u}^t$, obtaining for $E(\mathbf{u}\mathbf{u}^t|X_n)$ a diagonal matrix all of whose diagonal matrix elements are the same. However, in the absence of both the linear form of the conditional expectation and homoscedasticity, the form of the asymptotic SEs is more complicated.

Denote by J the 6×6 matrix $E[(1/n) X_n^t \mathbf{u}\mathbf{u}^t X_n]$. The entries in J do not depend on n and are, on the main diagonal, $E(u^2)$, $E(\delta^2 u^2)$, $E(\gamma^2 u^2)$, $E(x^2 u^2)$, $E(\delta^2 x^2 u^2)$, and $E(\gamma^2 x^2 u^2)$. All off-diagonal terms can be simplified in the same manner. Consider, for example, the entry in the fourth row and second column of J. This entry is $j_{42} = E[(1/n)\sum_{j=1}^{n} X_j u_j \sum_{i=1}^{n} w_i \delta_i]$. The independence of the different sibling pairs in the sample reduces this entry to $j_{42} = E[(1/n)\sum_{j=1}^{n} \delta_j x_j u_j^2] = E(\delta x u^2)$. Here, δ_j, x_j, and u_j are values for the jth sibling pair. Similarly, $j_{45} = E(\delta x^2 u^2)$, $j_{23} = E(\delta\gamma u^2)$, ... The formula for the asymptotic SE is then given as result 2.

Result 2

$\lim_{n\to\infty} E[\sqrt{n}(\mathbf{b}-\hat{\mathbf{b}})\sqrt{n}(\mathbf{b}-\hat{\mathbf{b}})^t] = M^{-1} \lim_{n\to\infty} E[(1/n) X_n^t \mathbf{u}\mathbf{u}^t X_n] M^{-1} = M^{-1} J M^{-1}$. The entries in J are complicated to compute, and full exposition of them would take us beyond the main purpose of the present paper. This is why we used simulations. For example, to compute $E(x^2u^2)$ it is necessary to condition on the genotypes at the QTL and on ibd status at the QTL, multiply by the appropriate probabilities, and then sum over all ibd and genotype possibilities. Conditioning in this way reduces the calculations to the calculation of the expected value of the square of the product of two random variables having a bivariate normal distribution. The formula is $E(x^2u^2|\text{ibd,genotypes}) = 2\text{cov}(x,u|\text{ibd,genotypes})^2 + E(x^2|\text{genotype})E(u^2|\text{genotype}) + 4E(x|\text{genotype})\,E(u|\text{genotype})\,\text{cov}(x,u|\text{ibd, genotypes})$.

In the selected-sample situation these general formulas remain valid. In the selected-sample model considered earlier all matrices are either 4×4 or $n \times 4$. The entries in J are more difficult to compute, but their form remains the same as in the unselected model.

Under the alternative hypothesis, simulations were used to avoid calculating the SEs for the elements for $\mathbf{b}$; however, under the null hypothesis, this was not done; in the latter case, a standard packaged least-squares statistical program was employed to calculate SEs. If the three normal distributions that constitute the phenotypic distribution are not widely separated, then the phenotypic distribution itself is very nearly normal. In this case, the least-squares program gives satisfactory results because of the near homoscedasticity and because of the resulting fact that J is approximately equal to $E(\mathbf{u}^2)$ times the matrix M. However, if the normal components are widely separated, then the least-squares package produces SE estimates that are too small. When the least-squares package is employed in this situation, a practical upper bound on the variance of the estimator divided by the least-squares estimate of its SE is about 1.20. This means that defining the critical region by using $1.645\sqrt{1.20} = 1.80$, rather than by using 1.645, will produce a type 1 error probability of at most .05 and, hence, an error on the side of caution.

While we recommend that investigators take this cautious approach when working with selected samples, our results concerning sample sizes and power are little affected when such an alteration of the critical region is introduced. Slightly shrinking the size of the critical regions for both selected and unselected sam-

ples has very little effect on the sample size and power of one relative to the other.

References

Amos CI, Elston RC (1989) Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genet Epidemiol 6:349–360

Amos CI, Elston RC, Wilson AF, Bailey-Wilson JE (1989) A more powerful robust sib-pair test of linkage for quantitative traits. Genet Epidemiol 6:435–449

Bishop DT, Williamson JA (1990) The power of identity-by-state methods for linkage analysis. Am J Hum Genet 46:254–265

Blackwelder WC, Elston RC (1982) Power and robustness of sib-pair linkage tests and extension to larger sibships. Commun Stat Theory Methods 11:449–484

Boehnke M, Moll PP (1989) Identifying pedigrees segregating at a major locus for a quantitative trait: an efficient strategy for linkage analysis. Am J Hum Genet 44:216–224

Cockerham CC, Weir BS (1983) Linkage between a marker locus and a quantitative trait of sibs. Am J Hum Genet 35:263–273

DeFries JC, Fulker DW (1988) Multiple regression analysis of twin data: etiology of deviant scores versus individual differences. Acta Genet Med Gemellol 37:205–216

Demenais FM, Amos CI (1989) Power of the sib-pair and lod-score methods for linkage analysis of quantitative traits. In: Elston RC, Spence MA, Hodge SE, MacCluer JW (eds) Multipoint mapping and linkage based upon affected pedigree members: Genetic Analysis Workshop 6. Alan R Liss, New York, pp 201–206

Goldgar DE (1990) Multipoint analysis of human quantitative genetic variation. Am J Hum Genet 47:957–967

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

Jayakar SD, Williamson JA, Zonta-Sgaramella L (1984) A nonparametric and parametric version of a test for the detection of the presence of a major gene applicable on data for the complete nuclear family. Hum Genet 67:143–150

Lange K, Spence MA, Frank MB (1976) Application of the lod method to the detection of linkage between a quantitative trait and a qualitative marker: a simulation experiment. Am J Hum Genet 28:167–173

Nance, WE, Neale MC (1989) Partitioned twin analysis: a power study. Behav Genet 19:143–150

Rao DC, Wette R, Ewens WJ (1988) Multifactorial analysis of family data ascertained through truncation: a comparative evaluation of two methods of statistical inference. Am J Hum Genet 42:506–515

Smith CAB (1975) A non-parametric test for linkage with a quantitative character. Ann Hum Genet 38:451–460