Am. J. Hum. Genet. 66:1631–1641, 2000

A Unified Sampling Approach for Multipoint Analysis of Qualitative and Quantitative Traits in Sib Pairs

Kung-Yee Liang,[1] Chiung-Yu Huang,[1] and Terri H. Beaty[2]

Departments of [1]Biostatistics and [2]Epidemiology, School of Hygiene and Public Health, Johns Hopkins University, Baltimore

Recent advances in molecular biology have enhanced the opportunity to conduct multipoint mapping for complex diseases. Concurrently, one sees a growing interest in the use of quantitative traits in linkage studies. Here, we present a multipoint sib-pair approach to locate the map position (τ) of a trait locus that controls the observed phenotype (qualitative or quantitative), along with a measure of statistical uncertainty. This method builds on a parametric representation for the expected identical-by-descent statistic at an arbitrary locus, conditional on an event reflecting the sampling scheme, such as affected sib pairs, for qualitative traits, or extreme discordant (ED) sib pairs, for quantitative traits. Our results suggest that the variance about $\hat{\tau}$, the estimator of τ, can be reduced by as much as 60%–70% by reducing the length of intervals between markers by one half. For quantitative traits, we examine the precision gain (measured by the variance reduction in $\hat{\tau}$) by genotyping extremely concordant (EC) sib pairs and including them along with ED sib pairs in the statistical analysis. The precision gain depends heavily on the residual correlation of the quantitative trait for sib pairs but considerably less on the allele frequency and exact genetic mechanism. Since complex traits involve multiple loci and, hence, the residual correlation cannot be ignored, our finding strongly suggests that one should incorporate EC sib pairs along with ED sib pairs, in both design and analysis. Finally, we empirically establish a simple linear relationship between the magnitude of precision gain and the ratio of the number of ED pairs to the number of EC pairs. This relationship allows investigators to address issues of cost effectiveness that are due to the need for phenotyping and genotyping subjects.

Introduction

Human genetics often focuses on quantitative traits associated with chronic diseases—such as total serum cholesterol, which is associated with heart disease; IGE, which is associated with asthma; and blood pressure, which is associated with hypertension—in an attempt to better understand genetic forces that control pathogenesis to complex diseases. Even when one is dealing with intrinsically discrete phenotypes (affected vs. not affected), such as the major psychiatric disorders, there are often several associated phenotypes available, sometimes called "endophenotypes," which are continuous in scale. Generally, there is more information available for statistical analysis of quantitative traits because all individuals (not just the affected individuals) contribute information. These quantitative traits will be used more frequently for mapping of susceptibility genes for complex diseases, since understanding the genetic control of risk factors for major chronic diseases may offer important opportunities for intervention.

Likelihood-based methods for detection of linkage between observed genetic markers and unobserved genes controlling quantitative traits is well developed; for example, see the work of Kruglyak and Lander (1995) and Allison et al. (1999) and references therein. Its primary drawback, shared by similar likelihood approaches for qualitative traits, is that the final conclusions about gene location are highly sensitive to the correct specification of mode of inheritance, which is required for computation of conventional LOD scores. Nonparametric approaches, such as the sib-pair design proposed by Haseman and Elston (1972), greatly alleviate such concern, because fewer assumptions are required. Instead, the Haseman-Elston method simply regresses the squared difference of the quantitative trait between members of a sib pair on the estimated number of marker alleles shared identical by descent (IBD). One can use least-squares methods to test the hypothesis of no linkage between an unobserved quantitative-trait locus (QTL) and the observed marker associated with estimated IBD sharing, by testing the statistical significance of this regression coefficient. This simple but elegant approach has drawn a good deal of attention recently, in two respects. One of these respects regards extensions of this method to multipoint analysis, in

Received November 17, 1999; accepted for publication February 18, 2000; electronically published April 5, 2000.

Address for correspondence and reprints: Dr. Kung-Yee Liang, Department of Biostatistics, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205. Email: kyliang@jhsph.edu

0002-9297/2000/6605-0018$02.00

which data from multiple markers are considered simultaneously (e.g., see Goldgar 1990; Fulker and Cardon 1994). Unlike the original single-marker approach, this multipoint extension has the advantage of estimating the location of the unobserved QTL at least to an interval between adjacent markers. Another respect has to do with the sampling considerations. Specifically, rather than drawing a random sample of sib pairs for inference, one may use ascertainment of pedigrees through one sib with extreme trait values to increase statistical power for detection of linkage (e.g., see Carey and Williamson 1991; Fulker et al. 1991; Eaves and Meyer 1994). The hypothesis of no linkage can then be tested by regression of the trait value of the unselected sib on the estimated IBD sharing for a marker between the two sibs. This approach of using selected sib pairs to map QTL has recently been extended to interval mapping in which two flanking markers are used (Cardon and Fulker 1994).

This unified approach for multipoint mapping by use of selected samples of sib pairs may be scrutinized as follows. First, Risch and Zhang (1995, 1996) pointed out that the power to detect linkage by use of random or selected samples of sibs can be improved on, sometimes substantially, by restriction of the sampling to extremely discordant (ED) pairs, since sib pairs that involve intermediate trait values provide relatively little power. Gu et al. (1996), Gu and Rao (1997), and Zhao et al. (1997) further suggested that a combination of reasonable numbers of ED and extremely concordant (EC) pairs may be even more powerful and cost effective. Second, with the trait value of the unselected sib as the dependent variable, any attempt to estimate, through regression methods, the location of QTL by multipoint mapping raises the following concern. Generally, the conventional least-squares estimate for the regression coefficients in linear-regression models may be biased when the sampling probability of each unit depends on the value of the dependent variable (e.g., see Liang and Qin, in press). Given that the trait values of each sib pair are likely to be positively correlated, the unselected sibs with extreme values are more favorably sampled than they otherwise would be. Here, the regression coefficient becomes a complicated function of the true (but unobserved) location of the QTL, as well as of other sampling factors.

Although the ED or EC designs have been shown to be more powerful for detection of linkage, their extension to multipoint analysis is less well studied, especially for mapping of unobserved genes. From the sampling viewpoint, it is more natural to consider the number of alleles IBD at the marker locus as being the dependent variable and the event that reflects the sampling scheme as being the independent variables (e.g., see Risch and Zhang 1995). With this in mind, one can, for multipoint analysis, simply regard the numbers of alleles IBD at multiple marker loci as being the dependent variables and regress these on the location of the markers for each sib pair. One implication of this observation is that the recent method developed by Liang et al. (in press) for multipoint mapping of genes that control qualitative traits can be readily applied to genes that control quantitative traits. The main goal of the present article is to propose a unified sampling approach for multipoint mapping of genes for both qualitative and quantitative traits, the latter focusing on the ED and EC designs. Just as in the work of Liang et al. (in press), the analysis method is designed to estimate the map position of an unobserved susceptibility gene, along with a measure of statistical uncertainty, under the assumption of some preliminary evidence of linkage in the region.

Robustness of IBD in Qualitative and Quantitative Traits

Consider a chromosomal region of length T cM framed by M markers at loci $0 \leqslant t_1 < , \ldots , < t_M \leqslant T$. We assume that the region contains no more than one unobserved susceptibility gene at some unknown location τ. We further assume that all M markers have been genotyped for each of n pairs of siblings. Let Φ denote the event that reflects the sampling criterion under which sib pairs were selected. For qualitative traits, a common criterion is that both siblings are affected, known as the affected-sib-pair (ASP) design. For quantitative traits, the following three sampling criteria have received attention lately: ED (denoted as "Φ_1"), EC with high trait values (Φ_2), and EC with low trait values (Φ_3). Let $S(t)$ be the number of alleles shared IBD for a sib pair at an arbitrary marker locus t, $0 \leqslant t \leqslant T$. For a specified sampling criterion Φ, one can express the expected IBD sharing of a marker, at t, in terms of its map distance from the true location τ, as

$$\begin{aligned}\mu(t) = E[S(t)|\Phi] &= 1 + (1 - 2\theta_{t,\tau})^2\{E[S(\tau)|\Phi] - 1\} \\ &= 1 + (1 - 2\theta_{t,\tau})^2 C_\Phi \ , \end{aligned} \qquad (1)$$

where $\theta_{t,\tau}$ is the recombination fraction between marker t and the unobserved gene at location τ. The proof of expression (1) for qualitative traits has been given by Liang et al. (in press) and can be applied directly to any arbitrary sampling criterion Φ as well.

REMARK 1. Expression (1), for the expected IBD sharing, $\mu(t)$, is valid regardless of the underlying mode of inheritance. A major assumption needed is that no more than one QTL is linked to the region, even though multiple genes may influence the phenotypic trait. It is our speculation that, when this assumption is violated, the

peak for the fitted $\mu(t)$ curve would be broader than that in the single-locus situation.

REMARK 2. One important implication of expression (1) with regard to the estimation of the true location τ of the QTL in the region is that $\mu(t)$ is monotonic in $|t - \tau|$ and attains its maximum or minimum value in $E[S(\tau)|\Phi]$ at $t = \tau$. However, whether $E[S(\tau)|\Phi]$ represents the maximum (or minimum) value depends on whether $C_\Phi = E[S(\tau)|\Phi] - 1$ in expression (1)—that is, whether $C_\Phi = \Pr[S(\tau) = 2|\Phi] - \Pr[S(\tau) = 0|\Phi]$ is positive (or negative). For ASPs and EC sib pairs with high-trait-value (Φ_2) and with low-trait-value (Φ_3) designs, this C_Φ term is likely to be positive when a marker is linked to the trait locus. For the ED design, on the other hand, the probability that zero alleles shared IBD at τ will be observed—that is, $\Pr[S(\tau) = 0|\Phi]$—is far greater than the probablility that two alleles shared IBD will be observed; in this situation, C_Φ is likely to be negative instead. These observations pose a challenging question as to how one may estimate τ when both ED and EC sib pairs are sampled, and this will be addressed in the next section.

REMARK 3. As noted by Liang et al. (in press), the ability to estimate τ well depends critically on the magnitude of C_Φ: the smaller C_Φ is in magnitude, the flatter the surface of $\mu(t)$ across the mapped region, which makes it difficult to resolve τ on the basis of IBD sharing (see fig. 2 in Liang et al., in press). For quantitative traits, sib pairs whose trait values are intermediate are likely to produce, on average, similar numbers of pairs that share two and zero alleles IBD, respectively. Consequently, the corresponding C_Φ value is likely to be closer to 0 and, hence, to provide less statistical power to detect linkage or to map the QTL. This observation is consistent with the argument that the use of sib pairs with extreme values (ED and/or EC) may be optimal sampling strategies for mapping of QTLs in humans (Gu and Rao 1997; Zhao et al. 1997). Note also that the square of C_Φ appears as the denominator of the sample-size (number of sib pairs) formula, both for a single marker (Risch and Zhang 1995) and for multiple markers (Liang et al., in press). Indeed, the ratio of the C_Φ squares from any two designs provides an excellent approximation to the ratio of sample sizes needed; see Liang et al. (in press).

To elaborate on the point made in Remark 3, we consider the following bivariate model (e.g., Haseman and Elston 1972; Risch and Zhang 1995; Gu et al. 1996; Zhao et al. 1997): $x_{ij} = \mu + g_{ij} + e_{ij}$, $j = 1, 2$, $i = 1, \ldots, n$. Here, μ is the overall mean of the quantitative trait for sib 1 (x_{i1}) and sib 2 (x_{i2}) from the ith of n sampled sib pairs, and g_{ij} and e_{ij} represent unobserved genetic and environmental effects, respectively. For simplicity, we assume that the trait locus that determines g_{ij} has two alleles, A_1 and A_2, with frequencies p and $q = 1 - p$, respectively. Furthermore, the genotypic effect is given by

$$g_{ij} = \begin{cases} a & \text{if genotype} = A_1A_1 \\ d & \text{if genotype} = A_1A_2 \\ -a & \text{if genotype} = A_2A_2 \end{cases}.$$

Thus, in the special case of an additive model, one has $a = 1$ and $d = 0$, whereas, for a dominant model, one has $a = 1 = d$. Finally, we assume that the residual environment (e_{i1}, e_{i2}) is bivariately normally distributed with mean (0,0), variance 1, and between sib-pair correlation ρ.

Tables 1 and 2 show the C_Φ values, classified by decile model (additive vs. dominant), by allele frequency ($p = .2$ vs. .4), and for two values of the residual correlation ($\rho = .0$ vs. .4), the same parameter configurations used by Risch and Zhang (1995). For example, for an additive model with $p = .4$ and $\rho = .4$ (lower section of table 1), the C_Φ value for a sib pair whose trait values are both in the top-10th decile is .17. For the same model, the C_Φ for an ED sib pair—that is, a sib pair whose trait value for sib 1 (2) is in the 1st (10th) decile, is $-.53$. Note that the ratio of the squares of these two C_Φ's $[(-.53/.17)^2 = 9.7]$ is in good agreement with the ratio of the corresponding sample sizes needed (107/10), as reported in table 2 of the work of Risch and Zhang (1995). On the basis of the magnitude (in absolute values) of these C_Φ values alone, it is rather clear that ED sib pairs will be more informative for estimation of τ. Furthermore, greater efficiency in estimating τ is gained for higher allele frequencies at the trait locus and for a stronger degree of residual correlation. To examine how sensitive these conclusions are to the assumption of normality of errors, tables 3 and 4 present the C_Φ values when the residuals are assumed to follow a bivariate logistic distribution that is known to have "heavier tails" than does the normal distribution. Because of the constraint associated with the range of the ρ values for bivariate logistic distributions (Johnson and Kotz 1972, p. 294), we used $\rho = .0$ and .3 instead of .4. Qualitatively, one would draw similar conclusions about both the superiority of the ED design and the impact of modeling assumptions, such as the magnitude of allele frequency and residual correlation, even in the presence of nonnormality.

Statistical Inference for Locating τ

Consider the design in which n_1, n_2, and n_3 sibling pairs of ED, EC with high trait values, and EC with low trait values, respectively, are recruited for linkage studies. Here $n = n_1 + n_2 + n_3$ is the total number of sib pairs sampled. Define, for $i = 1$, 2, and 3, $C_i = E[S(\tau)|\Phi_i] - 1$ as three unknown "nuisance" parameters. They are nui-

Table 1

C_Φ Values for the Additive Single-Locus Model with a Bivariate Normal Distribution for Residual Environment

	Value for Decile									
Decile	1	2	3	4	5	6	7	8	9	10
					$p = .2$					
	*	*	*	**	**	***	**	*	−.11	−.22
		*	**	*	**	***	**	**	*	−.17
1	*		**	**	**	***	**	**	*	−.14
2	*	*		**	**	***	***	**	**	−.10
3	**	**	**		***	***	***	**	**	*
4	***	**	**	**		***	***	***	***	**
5	**	**	**	**	**		***	**	**	**
6	*	**	***	**	**	**		**	**	*
7	−.10	**	**	***	**	**	**		*	.11
8	.16	*	*	**	***	**	**	**		.20
9	−.25	−.17	−.12	*	**	***	**	*	*	
10	−.40	−.31	−.25	−.20	−.15	−.10	**	***	*	.16
					$p = .4$					
	.16	.12	*	**	**	**	*	−.13	−.20	−.32
1	.13		**	**	**	***	**	*	*	−.16
2	*	*		**	**	***	***	**	*	−.11
3	**	*	*		**	**	***	***	**	*
4	**	**	*	*		**	**	**	***	**
5	*	***	**	**	**		**	**	**	**
6	−.13	**	***	**	**	**		*	*	*
7	−.20	**	**	***	**	**	*		*	.13
8	−.27	−.16	*	**	***	**	*	*		.21
9	−.37	−.25	−.17	−.10	*	***	**	*	.10	
10	−.53	−.40	−.31	−.23	−.17	−.10	**	**	*	.17

Note.—Above the diagonal, $\rho = .0$; below the diagonal, $\rho = .4$. * = $.05 \leq |C_\Phi| < .1$; ** = $.01 \leq |C_\Phi| < .05$; *** = $|C_\Phi| < .01$

sance in the sense that, in contrast to τ (the parameter of primary interest), these C_i's are needed only to completely specify $E[S(t)|\Phi_i]$ and are of little intrinsic interest in the mapping of the QTL. This is not to say that the magnitude of these nuisance parameters (the C_i's) have no bearing on statistical power for linkage analysis; see Remark 3. If all M markers are completely informative, so that IBD sharing can be counted directly, then, for each sib pair, one observes

$$S(Y_{ik}) = [S_{ik}(t_1), \ldots, S_{ik}(t_M)]' \quad \begin{matrix} k = 1, \ldots, n_i \\ i = 1,2,3 \end{matrix} \quad , \qquad (2)$$

where Y_{ik} denotes the marker information for the kth sib pair ascertained from the ith design, $i = 1$, 2, and 3, and $S_{ik}(t_j)$ denotes the number alleles shared IBD for the jth marker, $j = 1, \ldots, M$. In the more realistic situation, in which the markers are not necessarily fully informative, one may impute $S(t)$ by considering (e.g., see Liang et al., in press)

$$S(t_j|Y_{ik}) = \sum_{\ell=0}^{2} \ell \Pr\left(S_{ik}(t_j) = \ell | Y_{ik}\right) \ . \qquad (3)$$

Following the proof of Proposition 2 in the study by Liang et al. (in press), we can state that, for an arbitrary sampling criterion that includes Φ_1, Φ_2, and Φ_3, the expected value of the imputed IBD statistic is

$$E[S(t_j|Y_{ik})|\Phi_i] = 1 + (1 - 2\theta_{t,\tau})^2 C_i \equiv \mu_i(t_j) \ ; \qquad (4)$$

that is, the imputed statistic, $S(t_j|Y_{ik})$, in expression (3) is an unbiased estimator for $\mu_i(t_j)$, the expected IBD sharing, for any arbitrary marker at locus t_j, among those who are sampled through one or another sampling criterion Φ_i.

Remark 4. Expression (4) sheds light on how one may combine data from different sampling designs together to make inferences about τ, the location of a gene (see Remark 2). Analogous to the more familiar analysis-of-covariance model, expression (4) implies that, whereas the expected IBD sharing at an arbitrary marker locus may be different in magnitude under dif-

Table 2

C_Φ Values for the Dominant Single-Locus Model with a Bivariate Normal Distribution for Residual Environment

	Value for Decile									
Decile	1	2	3	4	5	6	7	8	9	10
					$p = .2$					
	.11	.11	.10	*	*	**	**	−.17	−.32	−.42
		.10	.10	*	*	**	**	−.16	−.30	−.40
1	.11		*	*	*	**	**	−.15	−.27	−.35
2	.10	.10		*	*	**	**	−.12	−.22	−.28
3	*	.10	.10		**	**	**	*	−.14	−.18
4	*	*	*	*		***	**	**	**	*
5	**	**	*	*	*		**	**	*	*
6	−.11	**	**	**	*	*		.10	.14	.16
7	−.25	−.14	*	**	**	*	*		.19	.21
8	−.37	−.29	−.22	−.14	*	**	*	.14		.23
9	−.44	−.40	−.35	−.29	−.19	*	*	.15	.19	
10	−.46	−.45	−.43	−.40	−.34	−.24	*	*	.18	.22
					$p = .4$					
	.24	.21	.16	*	*	−.18	−.27	−.34	−.38	−.40
		.19	.14	*	**	−.14	−.21	−.26	−.29	−.30
1	.23		.10	**	**	*	−.12	−.15	−.16	−.17
2	.18	.20		**	***	**	**	**	**	**
3	*	.14	.14		***	**	**	**	**	**
4	*	*	.10	.10		**	*	*	*	*
5	−.23	*	**	*	*		*	*	*	*
6	−.33	−.19	*	**	*	*		*	.10	.10
7	−.38	−.28	−.14	**	**	*	*		.10	.11
8	−.41	.34	−.21	*	**	*	*	.10		.11
9	−.43	−.38	−.28	−.14	**	**	*	.10	.10	
10	−.44	−.41	−.35	−.24	−.11	**	*	*	.10	.11

Note.—Above the diagonal, $\rho = .0$; below the diagonal, $\rho = .4$. * = $.05 \leq |C_\Phi| < .1$; ** = $.01 \leq |C_\Phi| < .05$; *** = $|C_\Phi| < .01$.

ferent sampling criteria, as characterized by these C_i's, the "effect" that the distance from a marker at locus t to the true location τ, as characterized by $\theta_{t,\tau}$, has on the expected IBD sharing is the same—namely, $(1 - 2\theta_{t,\tau})^2$. Thus, one is in position to estimate τ from the combined data, so long as (i) there is preliminary evidence of linkage to the region covered by the available map and (ii) there are at least two markers at different loci from this same region genotyped for each sib pair.

Define $S^*(Y_{ik})$ as

$$S^*(Y_{ik}) = [S(t_1|Y_{ik}), \ldots, S(t_M|Y_{ik})]' \quad , \qquad (5)$$

the imputed IBD-sharing statistic defined in expression (3) for the k sib pair obtained under the ith sampling criterion. We propose to estimate the vector $\delta = (\tau, C_1, C_2, C_3)'$ by solving the following estimating equation for δ:

$$\sum_{i=1}^{3} \sum_{k=1}^{n_i} \left[\frac{\partial \mu_i(\delta)}{\partial \delta}\right]' \mathrm{Cov}^{-1}[S^*(Y_{ik})][S^*(Y_{ik}) - \mu_i(\delta)] = 0 \quad , \qquad (6)$$

where

$$\mu_i(\delta) = [\mu_i(t_1; \tau, C_i), \ldots, \mu_i(t_M; \tau, C_i)]' \quad .$$

Here we have stressed the dependence of $\mu_i(t_j)$, $j = 1, \ldots, M$ on δ through the true location τ and C_i only, by reexpressing it as $\mu_i(t_j; \tau, C_i)$. Let $\hat{\delta}$ be the solution of the estimating equations in expression (6), and one has the following proposition.

Proposition 1. Let δ_0 be the true but unknown value for δ, and let λ_i be the limiting proportion of the total sample obtained under the ith sampling criterion, $i = 1$, 2, and 3—that is, $\lambda_i = \lim_{n \to \infty} n_i/n$, $i = 1$, 2, and 3. Under the assumption that there are reasonable numbers of sib pairs obtained under each of the three sampling criteria—that is, $0 < \lambda_i < 1$, $i = 1$, 2, and 3, then, as n becomes large, $\hat{\delta}$ is multivariate-normally distributed with mean δ_0 and a 4×4 covariance matrix:

$$\Sigma = \left\{ \sum_{i=1}^{3} \lambda_i \left[\frac{\partial \mu_i(\delta)}{\partial \delta}\right]' \mathrm{Cov}^{-1}[S^*(Y_i)] \left[\frac{\partial \mu_i(\delta)}{\partial \delta}\right] \right\}^{-1} \quad , \qquad (7)$$

evaluated at $\delta \equiv \delta_0$. Here, $S^*(Y_i)$ is the imputed IBD-sharing statistic defined in expression (5), for a sib pair obtained under the ith sampling criteria.

Table 3

C_Φ Values for the Additive Single-Locus Model with a Bivariate Logistic Distribution for Residual Environment

	Value for Decile									
Decile	1	2	3	4	5	6	7	8	9	10
					$p = .2$					
	*	*	*	**	**	***	**	**	−.12	−.19
		*	*	**	**	***	**	*	−.10	−.17
1	*		**	**	**	***	**	*	*	−.15
2	*	*		**	**	***	***	**	*	−.12
3	**	**	**		**	***	***	**	**	*
4	**	**	**	**		***	***	***	**	**
5	**	***	**	**	**		**	**	**	**
6	*	**	***	**	**	**		**	*	*
7	−.10	*	**	***	**	**	**		*	.12
8	−.16	−.10	*	**	***	**	**	*		.20
9	−.22	−.16	−.11	*	*	**	**	*	*	
10	−.31	−.25	−.21	−18	−.15	−.10	**	**	*	.18
					$p = .4$					
	.14	.12	*	*	***	**	*	−.14	−.20	−.27
		.11	*	**	**	**	*	−.11	−.16	−.23
1	.14		*	**	**	**	**	*	−.12	−.18
2	.11	*		**	**	***	**	**	*	−.12
3	*	*	*		**	**	**	***	**	*
4	**	**	*	*		**	**	**	***	**
5	*	***	**	*	*		**	**	**	**
6	−.12	**	***	**	*	*		*	*	*
7	−.17	−.10	**	***	**	*	*		.11	.14
8	−.24	−.16	−.10	*	***	**	*	*		.20
9	−.33	−.24	−.18	−.12	*	***	**	*	.10	
10	−.41	−.34	−.28	−.22	−.15	*	**	**	.11	.19

Note.—Above the diagonal, $\rho = .0$; below the diagonal, $\rho = .3$. * = $.05 \leq |C_\Phi| < .1$; ** = $.01 \leq C_\Phi| < .05$; *** = $|C_\Phi| < .01$

Remark 5. This generalized estimating equations (GEE) procedure, utilized by Liang et al. (in press) for the ASP design, was originally developed in the context of longitudinal data analysis (Liang and Zeger 1986). This method allows one to assess the precision of $\hat{\delta}$ even if the covariance matrix for $S^*(Y_{ik})$ in expression (6) is incorrectly specified. Specifically, one can estimate the covariance matrix Σ in espression (7) by $\hat{\Sigma} = \hat{\Sigma}_1^{-1}\hat{\Sigma}_2\hat{\Sigma}_1^{-1}$, where

$$\hat{\Sigma}_1 = \sum_{i=1}^{3}\sum_{k=1}^{n_i}\left[\frac{\partial \mu_i(\delta)}{\partial \delta}\right]' \mathrm{Cov}^{-1}[S^*(Y_{ik})]\left[\frac{\partial \mu_i(\delta)}{\partial \delta}\right],$$

$$\hat{\Sigma}_2 = \sum_{i=1}^{3}\sum_{k=1}^{n_i}\left[\frac{\partial \mu_i(\delta)}{\partial \delta}\right]' \mathrm{Cov}^{-1}[S^*(Y_{ik})] \times [S^*(Y_{ik}) - \mu_i(\delta)][S^*(Y_{ik}) - \mu_i(\delta)]' \mathrm{Cov}^{-1}[S^*(Y_{ik})]\frac{\partial \mu_i(\delta)}{\partial \delta},$$

evaluated at $\delta \equiv \hat{\delta}$.

Remark 6. As noted by Liang et al. (in press), a minor modification is needed when expression (6) is employed to locate τ. Strictly considered, $\theta_{t,\tau}$ is the genetic distance between the marker t and the true QTL—that is,

$$\theta_{t,\tau} = \frac{(1 - e^{-0.02|t-\tau|})}{2} \qquad (8)$$

under Haldane's (1919) mapping function. Note that this function is not differentiable with respect to τ. However, equation (8) can be modified by replacement of $|t - \tau|$ by

$$\begin{cases} |t-\tau| & \text{if } |t-\tau| \geq \epsilon \\ \dfrac{(t-\tau)^2}{2\epsilon} + \dfrac{\epsilon}{2} & \text{if } |t-\tau| < \epsilon \end{cases},$$

where ϵ is some prespecified positive number. Liang et al. (in press) found through simulations that this modified GEE method is insensitive to the choice of ϵ values. We will therefore use $\epsilon \equiv 1$ cM throughout the rest of the article.

We now study the variance of $\hat{\tau}$ for four designs for which M equally spaced and fully informative markers

Table 4

C_Φ Values for the Dominant Single-Locus Model with a Bivariate Logistic Distribution for Residual Environment

	Value for Decile									
Decile	1	2	3	4	5	6	7	8	9	10
					$p = .2$					
	.10	.10	.10	*	*	**	*	−.21	−.31	−.36
		.10	.10	*	*	**	*	−.20	−.30	−.35
1	.10		*	*	*	**	*	−.19	−.29	−.33
2	.10	.10		*	*	*	*	−.17	−.25	−.29
3	*	*	*		*	**	**	−.13	−.19	−.22
4	*	*	*	*		**	**	*	*	*
5	**	*	*	*	*		**	*	*	*
6	*	**	**	*	*	*		.13	.16	.17
7	−.25	−.16	*	**	**	*	.11		.19	.20
8	−.36	−.29	−.22	−.15	*	***	.10	.15		.22
9	−.40	−.35	−.31	−.27	−.21	−.10	**	.15	.19	
10	−.41	−.39	−.38	−.36	−.33	−.25	*	.12	.19	.21
					$p = .4$					
	.22	.21	.18	*	*	−.21	−.28	−.34	−.35	−.35
		.20	.16	*	*	−.19	−.24	−.27	−.29	−.30
1	.22		.13	*	*	−.13	−.17	−.19	−20	−.20
2	.20	.19		**	**	**	*	*	*	*
3	.13	.15	.16		**	**	**	**	**	**
4	*	**	.11	.11		*	*	*	*	*
5	−.24	−.10	***	*	*		*	*	*	*
6	−.32	−.20	*	**	*	*		*	*	.10
7	−.34	−.26	−.15	**	*	*	*		.10	.10
8	−.36	−.30	−.21	*	**	*	*	*		.10
9	−.37	−.33	−.28	−.15	**	***	*	*	.10	
10	−.39	−.38	−.35	−.24	*	**	*	*	.10	.10

Note.—Above the diagonal, $\rho = .0$; below the diagonal, $\rho = .3$. * = $.05 \leqslant |C_\Phi| < .01$; ** = $.01 \leqslant |C_\Phi| < .05$; *** = $|C_\Phi| < .01$.

are assumed to be genotyped for all sib pairs: design I, $M = 11$ (10 cM apart); design II, $M = 6$ (20 cM); design III, $M = 21$ (5 cM); and design IV, $M = 6$ (10 cM). Results are similar for designs in which markers are not equally spaced, and, therefore, such designs are not presented here. The first three designs all have the same length ($T = 100$ cM), framed by a varying number of markers with different map densities. Designs I and IV, on the other hand, have the same density, but the number of markers and total map length in design I are twice those of design IV. Figure 1 shows the ratio of $\text{var}(\hat{\tau})$ for designs II–IV versus design I. These patterns are the same irrespective of model of inheritance, allele frequency, and residual correlation. Results shown in figure 1 strongly suggest the benefit of having more dense markers available for genotyping, as far as the precision of the estimation of τ is concerned. The degree of efficiency gain depends, however, on the relative locations of τ, the true QTL, and the markers. When the $\text{var}(\hat{\tau})$ for design II (markers 20 cM apart) is compared with that of design I (markers 10 cM apart), for example, the range of the ratio is 1–3.3 (with 1 occurring when the trait locus τ occurs at an observable marker t_i). In other words, having markers 10 cM apart can reduce the $\text{var}(\hat{\tau})$ as much as 67%, compared with having markers 20 cM apart. The precision gain from having even denser markers (design III has markers 5 cM apart, compared with design I) is less striking yet still amounts to as much as a 60% reduction in $\text{var}(\hat{\tau})$. To further illuminate this phenomenon of efficiency gain, figure 2 plots $\text{var}(\hat{\tau})$ versus τ, $0 \leqslant \tau \leqslant 20$ cM, for designs I–III (under an additive model, $p = .2$, $\rho = .0$). For either design, figure 2 shows that the variance of $\hat{\tau}$ is at its minimum when τ, the true QTL, is located at the middle of flanking markers. Moreover, this minimum value of $\text{var}(\hat{\tau})$ can be attained by the addition of more markers between the original two flanking markers. Finally, one may argue that the reduction of $\text{var}(\hat{\tau})$ is a result of having more markers that are not necessarily dense. The variance ratio, as shown in figure 1, that is seen when design I (11 marker spaces, 10 cM apart) is contrasted with design IV (6 marker spaces, 10 cM apart), argues against this assertion. With the same marker density, there is no efficiency gain from having more markers,

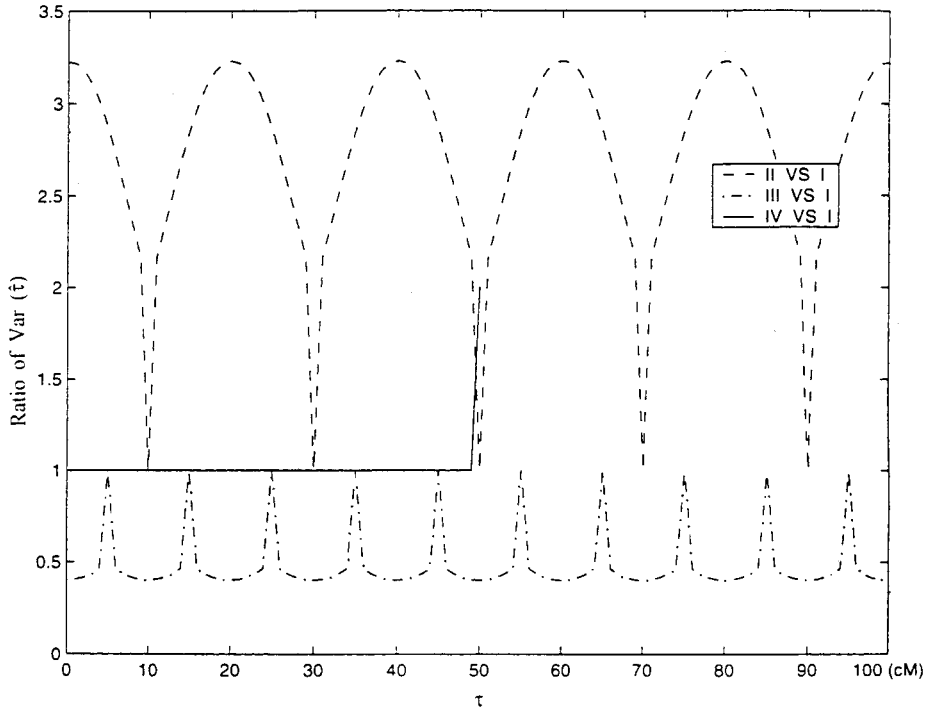

Figure 1 Ratio of asymptotic var($\hat{\tau}$) for designs II–IV versus design I for τ, the true QTL, $0 \leq \tau \leq 100$ cM

as long as the true QTL is flanked by the markers at hand.

In summary, when judged from the design viewpoint, our finding is consistent with the two-stage design commonly practiced today; that is, stage one is designed to identify a few regions in the genome that give preliminary evidence of linkage. Then, in stage two, more markers are added to the region, which gives some preliminary evidence of linkage. Our findings argue that this second stage will provide a more precise estimation of the location of a true unobserved gene—that is, will minimize the var($\hat{\tau}$).

Cost Effectiveness

Noting that Σ in equation (7) depends on the λ_i's, where $\lambda_1 + \lambda_2 + \lambda_3 = 1$, we can now examine the question as to how much efficiency, in terms of estimation of τ, can be gained by genotyping the EC sib pairs and including them along with ED sib pairs in the statistical analysis. Table 5 shows values of V_1/V_2 for some selected values of allele frequency and residual correlation. Here, V_1 is the asymptotic variance of $\hat{\tau}$ when the GEE method is used in equation (6), and V_2 is that of $\hat{\tau}$ when only the ED pairs were used—that is, when λ_2 and λ_3 are set to 0. Note that, with the bivariate normal genetic model specified in the previous section, not only the values of the C_i's but those of the λ_i's are completely determined. Only results from design I are presented here, and the ratios are rather stable, irrespective of the true location τ.

For the additive model, the efficiency loss, as measured by the variance of $\hat{\tau}$, will be between 46% and 53% (the end points of this range are for $p = .4$ and $p = .2$, respectively) if the residual correlation is 0 (i.e., $\rho = 0$). However, when $\rho = .4$, a much greater degree of efficiency loss occurs if only ED pairs are genotyped. In this situation—that is, $\rho = .4$, a higher proportion of all sib pairs fall into the EC categories relative to the ED category than when $\rho = .0$. For example, with $p = .2$ and $\rho = .0$, $(\lambda_1,\lambda_2,\lambda_3) = (.33,.39,.28)$, whereas, with $p = .2$ and $\rho = .4$, $(\lambda_1,\lambda_2,\lambda_3) = (.05,.49,.46)$. For the dominant model, the impact that residual correlation has on relative efficiency is as striking as that for the additive model. It is interesting to note that, when ρ is kept at the same level, the loss in efficiency is greater for the dominant model, with the larger allele frequency ($p = .4$), and thus a reversal of the pattern is seen with the additive model.

A natural question to raise next is how one translates these findings to address the issue of cost effectiveness. It is clear from table 5 that one can improve on the precision of mapping an unobserved gene by genotyping both EC and ED sib pairs, at the expense of additional genotyping and phenotyping costs. The question is, how great is the cost? To address this issue, figure 3 shows

the plot of V_1/V_2 against $\log[\lambda_1/(1-\lambda_1)]$, by these two models of inheritance (additive and dominant), by allele frequency (.2 and .4) and by the residual correlation (.0, .1, .2, .3, and .4). Combinations of all these parameters result in 20 pairs of points. Despite the fact that each point represents a different combination of the underlying genetic model, allele frequency, and residual correlation, figure 3 shows an approximate linear relationship between the loss in efficiency (V_1/V_2) and the proportion of ED sib pairs. Indeed, a simple linear-regression model fitted to these data leads to a model of the form $V_1/V_2 = 0.671 + 0.115\log[\lambda_1/(1-\lambda_2)]$ with $R^2 = .68$; thus, for example, to achieve a 50% reduction in variance of $\hat{\tau}$ (i.e., $V_1/V_2 = 0.5$), approximately five times as many EC sib pairs as ED sib pairs $[\lambda_1/(1-\lambda_1) \approx 0.22]$ should be included in the study population. This means that the cost of genotyping (and phenotyping) would be increased by fivefold if both ED and EC sib pairs were recruited. On the other hand, to achieve a 30% variance reduction ($V_1/V_2 = 0.7$), only 0.78 times as many EC sib pairs are needed $[\lambda/(1-\lambda) \approx 1.28]$, which results in a more modest, 0.8-fold increase in the original cost of genotyping.

Table 5

V_1/V_2

Genetic Model and Allele Frequency	Residual Correlation	V_1/V_2[a]
Additive:		
.2	.0	.47
.2	.4	.36
.4	.0	.54
.4	.4	.38
Dominant:		
.2	.0	.65
.2	.4	.31
.4	.0	.63
.4	.4	.29

[a] V_1 = the asymptotic variance of $\hat{\tau}$ when the GEE method is used; V_2 = the asymptotic variance of $\hat{\tau}$ when only the ED sib pairs are used.

Discussion

The allele-sharing method that uses sib pairs remains one of the most commonly used tools for linkage analysis. It provides a simple means for detection of linkage of a single marker (through hypothesis testing), without the need to specify the mode of inheritance. A drawback of this single-marker approach, which is not shared by the likelihood-based linkage approach, is its inability to estimate directly the genetic distance between the unobserved trait locus and the marker (as measured by θ), in the absence of the knowledge of the specific genetic mechanism. In recent years, however, the availability and

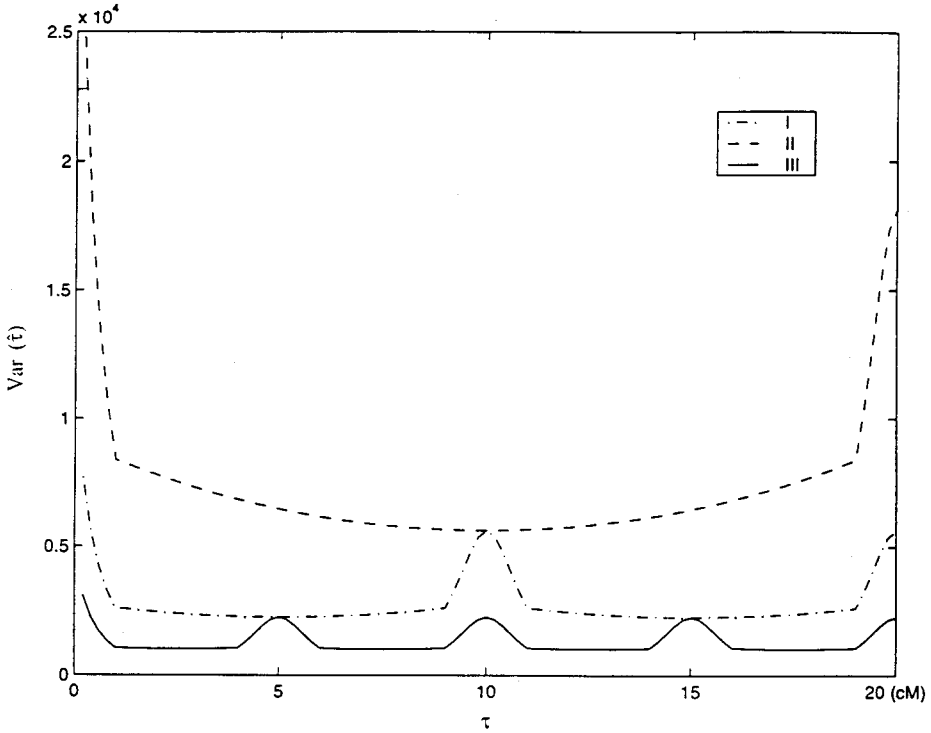

Figure 2 Asymptotic variance of $\hat{\tau}$ versus τ, the true QTL, $0 \leq \tau \leq 20$ cM for designs I–III. The assumed model is an additive model, with $p = .2$ and $\rho = .0$.

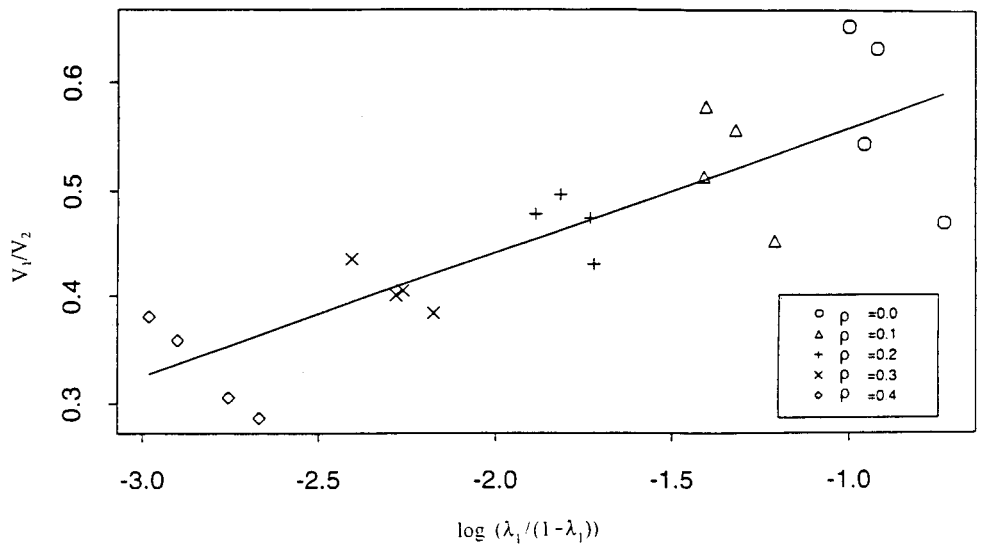

Figure 3 V_1/V_2 ratio versus $\log[\lambda_1/(1-\lambda_1)]$. Here, λ_1 is the proportion of ED sib pairs among ED and EC sib pairs

simultaneous use of multiple markers for linkage analysis has become commonplace. Concurrently, one sees a growing interest in the use of quantitative traits in linkage studies. A natural question arises as to whether a unified allele-sharing method for multipoint mapping of both qualitative and quantitative traits can be developed. The work by Kruglyak and Lander (1995) represents one of the first attempts to address this question.

In this article we have offered an alternative approach to the above question. Our approach capitalizes on the findings by Risch and Zhang (1995), Gu et al. (1996), and Zhao et al. (1997), noting that ED and EC sib pairs provide relatively more information for detection of linkage to an unobserved QTL, compared with unselected sib pairs. This observation allows one to divert the attention from a bivariate continuous distribution to the sampling of three discrete groups: ED, EC with high trait values, and EC with low trait values. These three groups are analogous to sib pairs with zero, one, and two affecteds, respectively, used in the analysis of qualitative traits. Furthermore, in accordance with the observation by Risch and Zhang (1995) that it is more natural to consider the IBD statistics as the dependent variable, we have presented, in expression (1), the expected IBD, at a marker locus, that is conditional on the event reflecting the sampling scheme, such as ED sib pairs for quantitative traits or ASPs for qualitative traits. In this representation, the relationship between the expected IBD at a marker locus and its genetic distance from the unobserved trait locus can still be established, without the need to specify the genetic mechanism. With this representation, one can carry out a unified parametric inference that uses, for example, the GEE method to locate an unobserved trait locus along with a measure of statistical uncertainty (i.e., confidence intervals) that uses either qualitative or quantitative phenotypes. As a side remark, the method presented here can be extended to either three or more siblings or to relative pairs other than siblings (see Liang et al., in press).

In addition, in this article we have addressed the issue as to how much efficiency is gained by incorporation of data from EC pairs when one is dealing with quantitative traits. Our results suggest that the gain in statistical efficiency, as measured by the reduction in $\mathrm{Var}(\hat{\tau})$, depends strongly on the magnitude of the residual correlation and, to a lesser extent, on the allele frequency and genetic mechanism (additive vs. dominant). Given that complex traits involve multiple genes—that is, that the residual correlation is likely to be large and cannot be ignored—one should routinely incorporate EC sib pairs in conjunction with ED sib pairs. Further work reveals that this reduction in variance can be explained effectively by a simple regression of the ratio of variances on $\log[\lambda_1/(1-\lambda_1)]$ as the independent variable. This simple linear relationship between the gain in statistical efficiency and the number of additional EC pairs needed allows one to address practical questions about the balance of cost effectiveness for the phenotyping and genotyping of subjects. Although this issue should be assessed on a case-by-case basis, recent work by Gu and Rao (1997) provides practical guidance to obtain an optimum design.

Finally, we will provide brief guidance for practition-

ers who are interested in applying the proposed work to the linkage data. As mentioned earlier, the proposed method for estimation of the location of susceptibility genes should be viewed as a supplement to the existing methods (e.g., LOD-score and NPL methods in GENEHUNTER), which are designed to detect linkage—that is, to test the null hypothesis of no linkage to the targeted region. Thus, when preliminary evidence of linkage within the chromosomal region is indicated, the method proposed here can be used to estimate the map position of the susceptibility gene, along with its confidence intervals. Such preliminary evidence could be obtained by performing either the LOD-score or the NPL method to test the null hypothesis of no linkage (e.g., $P < .01$). It is also helpful to explore map data in advance, by plotting of the imputed IBD statistics ($\bar{S}(t)$ in Liang et al., in press) against t in the region of interest. We are in the process of expanding the GENEFINDER program (Liang et al., in press) to accommodate multiple C's for quantitative traits with ED and EC designs. This program relies heavily on the access of investigators to other programs such as GENEHUNTER to (1) compute IBD statistics at each marker and (2) acquire preliminary evidence of linkage, and it will be available through the Web site, when properly documented and tested.

Acknowledgment

This work was supported by National Institutes of Health grant GM49909.

References

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

Cardon LR, Fulker DW (1994) The power of interval mapping of quantitative trait loci, using selected sib pairs. Am J Hum Genet 55:825–833

Carey G, Williamson J (1991) Linkage analysis of quantitative trails: increased power by using selected samples. Am J Hum Genet 49:786–796

Eaves L, Meyer J (1994) Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. Behav Genet 24:443–455

Fulker DW, Cardon LR (1994) A sib-pair approach to interval mapping of quantitative trait loci. Am J Hum Genet 54: 1092–1103

Fulker DW, Cardon LR, DeFries JC, Kimberling WJ, Pennington BF, Smith SD (1991) Multiple regression analysis of sib-pair data reading to detect quantitative trait loci. Reading Writing Interdisciplinary J 3:299–313

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

Gu C, Rao DC (1997) A linkage strategy for detection of human quantitative-trait loci. II. Optimization of study designs based on extreme sib pairs and generalized relative risk ratios. Am J Hum Genet 61:211–222

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

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309

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

Johnson NL, Kotz S (1972) Distributions in statistics: continuous multivariate distributions. John Wiley & Sons, New York

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

Liang KY, Chiu YF, Beaty TH. A robust identity by descent procedure using affected sib pairs: a multipoint approach for complex diseases. Hum Hered (in press)

Liang KY, Qin J. Regression analysis under non-standard situations: a pairwise pseudo-likelihood approach. J R Stat Soc Ser B (in press)

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

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

Zhao H, Zhang H, Rotter JI (1997) Cost-effective sib pair designs in the mapping of quantitative-trait loci. Am J Hum Genet 60:1211–1221