Using the Expectation or the Distribution of the Identity by Descent for Mapping Quantitative Trait Loci under the Random Model

Damian D. G. Gessler and Shizhong Xu

Department of Botany and Plant Sciences, University of California, Riverside, Riverside

Summary

We examine the ability of four implementations of the random model to map quantitative trait loci (QTLs). The implementations use either the expectation or the distribution of the identity-by-descent value at a putative QTL and either a 2 × 1 vector of sib-pair traits or their scalar difference. When the traits of both sibs are used, there is little difference between the expectation and distribution methods, while the expectation method suffers in both precision and power when the difference between traits is used. This is consistent with the prediction that the difference between the expectation and distribution methods is inversely proportional to the amount of information available for mapping. We find, though, that the amount of information must be very low for this difference to be noticeable. This is exemplified when both marker loci are fixed. In this case, while the expectation method is powerless to detect the QTL, the distribution method can still detect the presence (but not the position) of the QTL 59% of the time (when using trait values) or 14% of the time (when using trait differences). We also note a confounding between estimates of the QTL, polygenic, and error variance. The degree of confounding is small when the vector of trait values is used but can be substantial when the expectation method and trait differences are used. We discuss this in light of the general ability of the random model to partition these components.

Introduction

There is considerable interest in mapping quantitative trait loci (QTLs). Because the contributions of distinct quantitative loci cannot be directly observed, the mapping of QTLs is done by inferring their presence from the correlation between linked marker loci and individuals' measures for the quantitative traits (Sax 1923; Haseman and Elston 1972; Lander and Botstein 1989). Most QTL mapping has proceeded by using variations of this technique under the framework of the general linear model (GLM).

The GLM can test either fixed or random effects, depending on whether the inference space corresponds to only and exactly the effects tested or to a universe of all possible effects. While it may seem desirable in QTL mapping always to use the expanded inference space of the random model, the necessary use of inbred lines in much QTL methodology has dictated a fixed model approach. Both linear regression (Haley and Knott 1992; Martínez and Curnow 1992) and maximum-likelihood methods (Lander and Botstein 1989; Knott and Haley 1992) have been used to solve fixed model GLMs. In the regression method, a best-fit is achieved by minimizing the squared deviation between the predicted and observed values for the trait as a function of the mean probability that a given allele at a putative QTL is shared among sibs, given the observed marker information. In a full-sib model, this mean probability reflects a weighting of the presence or absence of a shared allele; this weighting, denoted $\hat{\pi}_i$ for the *i*th family, is an estimate of the mean identity-by-descent (IBD) value at the putative QTL. For example, if there is a 25% chance that two sibs share no alleles, a 50% chance that they share exactly one, and a 25% chance that they share two, then $\hat{\pi}_i = \frac{1}{4} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{4} \cdot 1 = \frac{1}{2}$.

Alternatively, the maximum-likelihood method does not use the expectation $\hat{\pi}_i$, but uses the distribution of π_i, i.e., it incorporates the discrete probabilities that sibs may share exactly zero, one, or two alleles at the putative QTL. Haley and Knott (1992) showed that in a fixed model both methods can produce similar results, though the regression method as defined above may overestimate the true residual error (Xu 1995).

If parents are arbitrarily outbred, then matings necessarily involve a sampling of alleles, and thus the relevant inference space presupposes the random model. Kruglyak and Lander (1995) used the distribution of π_i in a multipoint sib-pair analysis and noted its advantages in terms of the amount of information extracted and its adherence to a strict maximum-likelihood model. Yet it is unclear whether the advantages of this method will be large enough to warrant its exclusive adoption, or,

Received May 9, 1996; accepted for publication September 9, 1996.

Address for correspondence and reprints: Dr. Damian D. G. Gessler, Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA 93531. E-mail: gessler@evolution.ucr.edu

0002-9297/96/5906-0025$02.00

as in the fixed model case, most situations will yield similar results when either the expectation or the distribution of π_i is used. We refer to these two methods as the *expectation* and *distribution* methods. For the random model, both the expectation and the distribution method can be incorporated into a single maximum-likelihood model by an appropriately defined log-likelihood function.

In this paper, we examine the consequences of using the expectation or the distribution method under a general maximum-likelihood methodology. We do this for two models, one where we use traits from each of two full sibs and the other where we use only the difference between the trait values of each sib pair (Haseman and Elston 1972; Fulker and Cardon 1994). We choose to work in a parameter space that is particularly relevant to human genetics. For example, it is known that maximum-likelihood estimates often require large sample sizes and that the selection of discordant sibs and the use of inbred lines and can increase the power of QTL mapping (Paterson et al. 1988; Lander and Botstein 1989; Risch and Zhang 1995). In human populations, unknown pedigrees and relatively small family sizes can restrict the use of these options. We concentrate on using no more than 1,000 families, exactly two full-sibs per family with no selection, and we restrict ourselves to randomly constituted parental populations. It is important to note that we assume our data to consist of only a list of quantitative measures and fully expressive, codominant marker genotypes. We exclude the possibility of manipulative mating schemes or pedigree analysis.

The Model

We consider a model similar to that described by Goldgar (1990) and Schork (1993):

$$y_{ij} = \mu + g_i + a_i + \varepsilon_{ij} , \qquad (1)$$

where y_{ij} is the observed effect (the measured value of the trait) for the jth sib ($j = 1, 2$) of the ith family ($i = 1, 2, \ldots n$), μ is the grand mean, g_i is the contribution at the putative QTL, a_i is the contribution of all other (and presumably unlinked) QTLs, and ε_{ij} is the error term. We will sometimes refer to a_i as the "polygenic contribution." In the formulation of (1), the parental populations of individual y_{ij} may be arbitrarily outbred, and thus (1) is treated as a random model. Because of this, testing for the significance of effects g_i and a_i in (1) is equivalent to testing for significant variance components:

$$\mathrm{var}(Y) = \sigma^2 = \sigma_g^2 + \sigma_a^2 + \sigma_\varepsilon^2 ,$$

where Y is a random variable for y_{ij}. We follow the standard technique of Lander and Botstein (1989) and test successive putative QTLs against the null hypothesis of no QTL, i.e., $y_{ij} = \mu + a_i + \varepsilon_{ij}$.

To test model (1) we construct the bivariate normal probability density function

$$f(\mathbf{y}_i) = \frac{1}{2\Pi\sigma^2|\mathbf{C}_i|^{1/2}} \times \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_i - \mathbf{1}\mu)'\mathbf{C}_i^{-1}(\mathbf{y}_i - \mathbf{1}\mu)\right\} \qquad (2)$$

for observing effects $\mathbf{y}_i = [y_{i1}, y_{i2}]'$ from family i (Xu and Atchley 1995). In the above equation $\mathbf{1} = [1,1]'$ and $\mathbf{C}_i$ is defined such that

$$\mathrm{var}(\mathbf{y}_i) = \sigma^2\mathbf{C}_i = \sigma^2\begin{bmatrix} 1 & r_i \\ r_i & 1 \end{bmatrix} ,$$

where $r_i = \tilde{\pi}_i h_g^2 + {}^1\!/_2 h_a^2$, $h_g^2 = \sigma_g^2/\sigma^2$, and $h_a^2 = \sigma_a^2/\sigma^2$. To model the expectation method, we replace $\tilde{\pi}_i$ (the IBD value at the putative QTL) by its expectation $\tilde{\pi} = \bar{\pi}_i = 0p_0 + {}^1\!/_2 p_{1/2} + 1p_1$, where p_0, $p_{1/2}$, and p_1 are the probabilities of sharing zero, one, or two alleles. For clarity, we omit the subscript i from p_0, $p_{1/2}$, and p_1 and present formulas for them in the next section. We then employ the usual method of solving for μ, σ^2, h_g^2, and h_a^2 by numerically estimating the log-likelihood function

$$L_1 = \sum_{i=1}^{n} \ln f(\mathbf{y}_i) . \qquad (3)$$

For the distribution method, we incorporate the distribution of π_i and define a new log-likelihood function as

$$L_2 = \sum_{i=1}^{n} \ln[p_0 f_0(\mathbf{y}_i) + p_{1/2} f_{1/2}(\mathbf{y}_i) + p_1 f_1(\mathbf{y}_i)] , \qquad (4)$$

where $f_x(\mathbf{y}_i)$ is $f(\mathbf{y}_i)$ evaluated at $x = \tilde{\pi}_i = 0, {}^1\!/_2$, or 1.

If there is no variance in $\tilde{\pi}_i$ then (3) and (4) are equal. This condition is approached in fixed-effect models with inbred lines, but it is rarely attained (i.e., it is only a limiting case) in outbred populations.

Simulation Methods

All simulations follow a similar procedure. We construct canonical alleles with constant frequencies from one common grandparent population. These alleles constitute all marker loci and the specific QTL we are interested in mapping. From this population, we generate four gametes, and from these gametes we construct two parents. Separate from the marker loci and QTL, in some simulations each parent segregates an additional 12 unlinked, biallelic loci as a polygenic contribution to the trait. The effect of each polygenic allele is distributed as a normal variate with mean zero and variance $\sigma_a^2/24$. From these parents, we generate two full sibs, each sib being the product of an independent meiotic event. We conducted runs using both randomly placed chiasmata and Haldane's (1919) mapping function to determine the probability of recombinants, but here report results using only the mapping-function technique.

We use the above algorithm to generate n families, with parents for each family being picked anew from the invariant infinite-size grandparent population. Using the mean and variance of the QTL effect in the grandparent population, we transform the QTL effect in each sib so that the population of sib effects is distributed normally with mean zero and variance $\sigma_g^2 \cong 12.5$. The total effect for each sib is the sum of this effect, its polygenic contribution, and any environmental effect. Environmental effects are distributed as normal variates with mean zero and variance σ_e^2. From these values, we generate a phenotype for each sib as $y_{ij} = g_{ij} + a_{ij} + e_{ij}$.

As each sib is made, we look at its and its parents' marker alleles and try to infer its genotype. To do this, we assume knowledge of the linkage phase of each individual. Knowledge of the linkage phase restricts the number of possible offspring types, and, by decreasing the variance in estimating each sib's IBD value, it increases the amount of information extracted. Since increased information reduces the variance in $\hat{\pi}_i$, assuming knowledge of the linkage phase is conservative with respect to identifying a difference between the expectation and distribution methods. Concurrently, though, the additional knowledge increases the chance of detecting a QTL if one is segregating at a site and therefore yields a procedure that is upwardly biased in detecting the presence of a QTL.

Formulas for the probability of IBD based on two flanking markers have been published elsewhere, and, using our notation,

$$p_0 = \Pr(\pi_i = 0) = [p_1^m(1 - p_2^m) + p_2^m(1 - p_1^m)] \times [p_1^f(1 - p_2^f) + p_2^f(1 - p_1^f)]$$

$$p_1 = \Pr(\pi_i = 1) = [p_1^m p_2^m + (1 - p_1^m)(1 - p_2^m)] \times [p_1^f p_2^f + (1 - p_1^f)(1 - p_2^f)]$$

$$p_{1/2} = \Pr(\pi_i = 1/2) = 1 - p_0 - p_1 ,$$

where p_y^x is the probability that an allele at the QTL comes from the mother's or father's ($x = m, f$) first or second homologue ($y = 1, 2$), conditional on the observed marker genotype. These probabilities are a function of the recombination distance between the marker loci and the QTL and are shown in table 1. If the parents' genotypes are $(A_1^m Q_1^m B_1^m)/(A_2^m Q_2^m B_2^m)$ and $(A_1^f Q_1^f B_1^f)/(A_2^f Q_2^f B_2^f)$ then $p_y^x = \Pr(Q_y^x \mid \text{marker genotype})$. With arbitrarily outbred parental populations, not all genotypes can be uniquely determined from the offspring's marker alleles, and one cannot unequivocally determine p_0, $p_{1/2}$, and p_1. In these cases, we use a weighted average for p_0, $p_{1/2}$, and p_1 taken over all possible parent-offspring combinations.

From the vector of trait values y, the above formulas, and table 1, we then maximize the log-likelihood functions (3) and (4), using an implementation of the simplex algorithm of Nelder and Mead (1965). For each putative position, we calculate the log-likelihood ratio LR = $-2(L_0 - L_A)$, where L_0 is the value of the log-likelihood function under the null model and L_A (A = {1,2}) is the log-likelihood function under the alternative model. The position with the highest LR is deemed the estimate of the position of the QTL. For some positions, the maximizing algorithm found spurious local maxima, as noted by a substantially negative LR. If this happened, we restarted the search with a different set of initial conditions; if this continued to happen more than five times for a single position, we abandoned the run. The likelihood of the numerical failure of the algorithm varied across different parameter combinations, but in no case did it constitute more than a few percentage points of all attempted runs.

To examine the differences between the expectation and distribution methods, we construct four parameter classes, each with three levels of resolution. With $h_a^2 = 0$, we examine (1) QTL heritabilities of $h_g^2 = 0.25$, 0.50, and 0.75; (2) low, medium, and high levels of marker informativeness; (3) population sizes of 250, 500, and 1,000 families; and (4) interval sizes of 10, 20, and 40 cM. For low marker informativeness, each marker locus in the grandparent population has two alleles at frequencies of 0.9 and 0.1, respectively. For medium marker informativeness, each has two alleles at equal frequency, and, for high informativeness, there are six alleles, all at equal frequency. In all cases, six equally frequent alleles segregate at the QTL. While testing each level, we hold all other parameters at their intermediate value. We examine additional cases with zero marker informativeness and/or nonzero polygenic heritability—the details of which we describe below. Each class of simulations is repeated 300 times.

To determine the critical test statistic for a particular parameter class, we performed 1,000 simulations each with no QTL segregating. We then created a list of the largest *LR* found from each run, and used the 50th (5%) largest value as the critical *LR*. When removing the QTL, we maintained all other variances constant, thus reducing the overall variance for the trait. Alternatively, we could have changed the environmental and/or poly-

Table 1

Conditional Probability of a QTL Allele, Given the Flanking Marker Genotype

Marker Offspring Genotype	4 × Probability	QTL Allele			
		$p_1^m = \Pr(Q_1^m \mid M)$	$p_2^m = \Pr(Q_2^m \mid M)$	$p_1^f = \Pr(Q_1^f \mid M)$	$p_2^f = \Pr(Q_2^f \mid M)$
$A_1^m A_1^f B_1^m B_1^f$	$(1 - r)^2$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A r_B/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A r_B/(1 - r)$
$A_1^m A_1^f B_1^m B_2^f$	$r(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A r_B/(1 - r)$	$(1 - r_A) r_B/r$	$r_A(1 - r_B)/r$
$A_1^m A_2^f B_1^m B_1^f$	$r(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A r_B/(1 - r)$	$r_A(1 - r_B)/r$	$(1 - r_A) r_B/r$
$A_1^m A_2^f B_1^m B_2^f$	$(1 - r)^2$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A r_B/(1 - r)$	$r_A r_B/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$
$A_1^m A_1^f B_2^m B_1^f$	$r(1 - r)$	$(1 - r_A) r_B/r$	$r_A(1 - r_B)/r$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A r_B/(1 - r)$
$A_1^m A_1^f B_2^m B_2^f$	r^2	$(1 - r_A) r_B/r$	$r_A(1 - r_B)/r$	$(1 - r_A) r_B/r$	$r_A(1 - r_B)/r$
$A_1^m A_2^f B_2^m B_1^f$	r^2	$(1 - r_A) r_B/r$	$r_A(1 - r_B)/r$	$r_A(1 - r_B)/r$	$(1 - r_A) r_B/r$
$A_1^m A_2^f B_2^m B_2^f$	$r(1 - r)$	$(1 - r_A) r_B/r$	$r_A(1 - r_B)/r$	$r_A r_B/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$
$A_2^m A_1^f B_1^m B_1^f$	$r(1 - r)$	$r_A(1 - r_B)/r$	$(1 - r_A) r_B/r$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A r_B/(1 - r)$
$A_2^m A_1^f B_1^m B_2^f$	r^2	$r_A(1 - r_B)/r$	$(1 - r_A) r_B/r$	$(1 - r_A) r_B/r$	$r_A(1 - r_B)/r$
$A_2^m A_2^f B_1^m B_1^f$	r^2	$r_A(1 - r_B)/r$	$(1 - r_A) r_B/r$	$r_A(1 - r_B)/r$	$(1 - r_A) r_B/r$
$A_2^m A_2^f B_1^m B_2^f$	$r(1 - r)$	$r_A(1 - r_B)/r$	$(1 - r_A) r_B/r$	$r_A r_B/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$
$A_2^m A_1^f B_2^m B_1^f$	$(1 - r)^2$	$r_A r_B/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A r_B/(1 - r)$
$A_2^m A_1^f B_2^m B_2^f$	$r(1 - r)$	$r_A r_B/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$	$(1 - r_A) r_B/r$	$r_A(1 - r_B)/r$
$A_2^m A_2^f B_2^m B_1^f$	$r(1 - r)$	$r_A r_B/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A(1 - r_B)/r$	$(1 - r_A) r_B/r$
$A_2^m A_2^f B_2^m B_2^f$	$(1 - r)^2$	$r_A r_B/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$	$r_A r_B/(1 - r)$	$(1 - r_A)(1 - r_B)/(1 - r)$

NOTE.—r is the recombination fraction between the two flanking markers A and B; r_A and r_B are the recombination fractions between each marker and the QTL. The marker probability is the relative likelihood of observing the marker genotype given the recombination distance between them.

genic variance to maintain the overall variance or genetic heritability. In terms of the critical *LR,* we found little difference between the two null hypotheses.

Finally, we differentiated between using a vector of sib-pair values and their difference by implementing one as the sib-pair model defined above and the other as a sib-pair–difference model. For the sib-pair–difference model, we used the one-dimensional normal probability density function

$$f(\tilde{y}_i) = \frac{1}{(2\Pi\tilde{\sigma}_i^2)^{1/2}} \exp\left\{-\frac{\tilde{y}_i^2}{2\tilde{\sigma}_i^2}\right\},$$

where $\tilde{y}_i = y_{i1} - y_{i2}$, $\tilde{\sigma}_i^2 + \sigma^2[2(1 - \tilde{\pi}_i)h_g^2 + 2h_e^2 + h_a^2]$, and $h_e^2 + \sigma_e^2/\sigma^2$. The maximum-likelihood functions for the expectation and distribution methods are applied analogously as before to $\tilde{\pi}_i$.

Results

Sib-Pair Model

Table 2 reports the estimated QTL position, the coefficient of variation for the position, the total phenotypic variance, and the heritabilities h_g^2 and h_a^2 for the first four parameter classes for the sib-pair model. In no case is there any significant difference between the expectation and distribution methods.

Some trends are evident in the table, although in general the effects are weak. In all cases, the position and total phenotypic variance are successfully predicted, while the observed estimates of $h_g^2 + h_a^2$ only slightly exceed their expectation. In general, the sums $h_g^2 + h_a^2 = 0.25, 0.50$, or 0.75 are conserved, implying a successful partitioning of the genetic and residual variances. Because the modeled polygenic heritability is zero, any partitioning into h_a^2 under a conserved sum tends to reduce h_g^2, and thus the QTL heritability estimates are biased low.

There is a strong effect of the level of heritability, informativeness, number of families, and interval size on decreasing the coefficient of variation (CV) in the estimate of the QTL position, although there is no difference between the expectation and distribution methods. In general, increasing the information content reduces the CV in the QTL position from ~0.8 to ~0.5.

The partitioning of $h_g^2 + h_a^2$ is correspondingly sensitive to the information content. A reduction in marker informativeness leads to an increase in the confounding of h_g^2 and h_a^2, although we do not observe a reciprocal increase at the higher level of informativeness. A similar effect is seen with population sizes, where 250 families show a greater confounding between h_g^2 and h_a^2 than do 500 families, while there is no significant benefit in increasing the size to 1,000. As expected, both increasing the informativeness and the family size reduces the variance of the various estimates. The preceding is inversely true for interval sizes, where increasing the interval to 40 cM significantly compromises the precision and abil-

Table 2

Estimates of the Position, Coefficient of Variation in the Position (CV), Total Phenotypic Variance (σ^2), and QTL (h_g^2) and Polygenic (h_a^2) Heritabilities for the Expectation and Distribution Methods

	Expectation Method					Distribution Method				
	Position	CV	σ^2	h_g^2	h_a^2	Position	CV	σ^2	h_g^2	h_a^2
Standard	10.25 (6.43)	.63	25.04 (1.21)	.43 (.1174)	.09 (.1220)	10.24 (6.57)	.64	25.04 (1.21)	.43 (.1141)	.09 (.1207)
$h_g^2 = .25$	10.66 (7.63)	.72	49.87 (2.42)	.21 (.1030)	.06 (.0969)	10.72 (7.61)	.71	49.87 (2.42)	.20 (.1055)	.07 (.0990)
$h_g^2 = .75$	10.28 (5.05)	.49	16.72 (.72)	.69 (.1094)	.08 (.1194)	10.28 (4.99)	.49	16.73 (.72)	.68 (.0963)	.09 (.1122)
Low	9.50 (7.46)	.79	24.92 (1.09)	.39 (.1641)	.12 (.1662)	9.57 (7.44)	.78	24.92 (1.09)	.38 (.1607)	.13 (.1606)
High	9.86 (5.17)	.52	25.02 (1.11)	.44 (.1047)	.08 (.1081)	9.91 (5.31)	.54	25.02 (1.12)	.44 (.1040)	.08 (.1090)
N = 250	10.25 (7.21)	.70	24.94 (1.55)	.41 (.1611)	.11 (.1605)	10.39 (7.34)	.71	24.94 (1.55)	.40 (.1609)	.11 (.1622)
N = 1,000	9.80 (5.40)	.55	24.96 (.75)	.44 (.0878)	.07 (.0886)	9.79 (5.59)	.57	24.97 (.75)	.44 (.0871)	.07 (.0895)
10 cM	5.23 (2.92)	.56	24.88 (1.13)	.45 (.1123)	.07 (.1084)	5.17 (2.95)	.57	24.88 (1.13)	.45 (.1113)	.07 (.1105)
40 cM	19.43 (13.68)	.70	24.96 (1.11)	.37 (.1516)	.12 (.1632)	19.54 (13.84)	.71	24.96 (1.11)	.37 (.1482)	.13 (.1611)

Note.—The standard run is $h_g^2 = .50$, medium marker information as defined in the text, N = 500 families, and a 20-cM marker interval. Each additional run differs from the standard run by the parameter change noted in the left-hand column. The true position of the QTL is in the middle of the interval. The expected total phenotypic variance is 50, 25, and 16.67 for h_g^2 = .25, .50, and .75, respectively. Standard errors are in parentheses.

ity to partition the heritabilities, while reducing the interval size to 10 cM has virtually no effect.

Sib-Pair–Difference Model

Table 3 reports analogous results for the sib-pair–difference model. Here, there is a difference between the expectation and distribution methods. Both methods under both models produce comparable estimates of the QTL position, but in the sib-pair–difference model the distribution method is more accurate in estimating the total phenotypic variance and partitioning h_g^2 and h_a^2. In fact, the distribution/sib-pair–difference combination produces the most accurate estimates of the polygenic heritability (i.e., consistently the closest estimates for $h_a^2 = 0$). The expectation method consistently produces upwardly biased estimates of the total phenotypic variance while failing to conserve the sum $h_g^2 + h_a^2$.

Power

With one important exception, there is no significant difference between the expectation and distribution methods when examining the power to detect the QTL. Table 4 reports the power in terms of the percentage of runs with a LR at least as large as the critical LR. The table includes a special class of runs with zero marker informativeness. It is in this class that the distribution method shows a substantial ability to detect the presence, though not the position, of the QTL.

In all except two cases, the sib-pair model is more powerful than the sib-pair–difference model in detecting

Table 3

Estimates of the Position, Coefficient of Variation in the Position (CV), Total Phenotypic Variance (σ^2), and QTL (h_g^2) and Polygenic (h_a^2) Heritabilities for the Expectation and Distribution Methods, Using the Sib-Pair Difference Model

	Expectation Model					Distribution Method				
	Position	CV	σ^2	h_g^2	h_a^2	Position	CV	σ^2	h_g^2	h_a^2
Standard	10.29 (6.48)	.63	26.48 (3.22)	.45 (.1552)	.12 (.1120)	10.31 (6.75)	.65	24.94 (2.75)	.45 (.1398)	.03 (.0188)
$h_g^2 = .25$	9.85 (7.50)	.76	57.27 (6.98)	.24 (.1627)	.22 (.1419)	9.78 (7.66)	.78	51.10 (6.08)	.26 (.1635)	.02 (.0081)
$h_g^2 = .75$	10.00 (4.98)	.50	17.93 (1.79)	.70 (.1311)	.13 (.1371)	9.92 (4.97)	.50	16.81 (1.66)	.71 (.0804)	.04 (.0160)
Low	9.62 (7.24)	.75	27.06 (4.14)	.47 (.2286)	.12 (.1216)	9.45 (7.57)	.80	24.76 (3.48)	.43 (.1866)	.04 (.0264)
High	10.14 (5.39)	.53	26.54 (3.17)	.47 (.1183)	.11 (.1144)	10.26 (5.60)	.55	25.27 (2.74)	.48 (.1098)	.03 (.0156)
N = 250	9.29 (7.08)	.76	26.93 (4.54)	.46 (.2151)	.12 (.1374)	9.33 (7.39)	.79	25.22 (4.12)	.45 (.1975)	.04 (.0495)
N = 1,000	10.31 (5.56)	.54	26.86 (3.00)	.46 (.1214)	.13 (.1178)	10.29 (5.69)	.55	25.17 (2.26)	.47 (.1075)	.03 (.0099)
10 cM	5.15 (2.97)	.58	26.70 (3.27)	.48 (.1247)	.11 (.1041)	5.09 (3.03)	.60	25.40 (2.86)	.48 (.1189)	.03 (.0128)
40 cM	17.99 (12.64)	.70	26.98 (3.91)	.44 (.2139)	.14 (.1382)	17.82 (13.37)	.75	24.99 (3.45)	.43 (.1829)	.04 (.0346)

Note.—All information as is in table 2.

Table 4

Power to Detect the QTL

	Method	
Model	Expectation	Distribution
Standard Parameters	.96	.96
	.67	.64
h_g^2 = .25	.62	.60
	.17	.15
h_g^2 = .75	1.00	1.00
	1.00	.997
Marker Informativeness:		
Zero	3.33×10^{-3} [a]	.59
	.05	.14
Low	.79	.81
	.40	.41
High	.98	.98
	.91	.91
N = 250	.86	.86
	.42	.43
N = 1,000	1.00	1.00
	.90	.92
10-cM interval	.97	.97
	.84	.85
40-cM interval	.82	.81
	.48	.50

Note.—Each line reports the proportion of runs that had at least one LR test statistic greater than the critical test statistic at the 95% level. For each set of simulations, the upper line is for the sib-pair model; the lower line is for the sib-pair difference model. The standard model is described in the text and table 1.

[a] Only 1 case in 300.

the QTL. Of the two exceptions, one is equivalent to a difference in the size of the observed type I error rate for the expectation method (zero marker informativeness) and the other is where both models detect the QTL in all 300 simulations (h_g^2 = 0.75).

Nonzero Polygenic Heritability

We ran three × 300 additional simulations to determine how some of the above conclusions apply when there is nonzero polygenic heritability. For these simulations we examined only the sib-pair model. The results are reported in tables 5 and 6. In general, there continues to be a relatively large variance around the estimate of the position of the QTL, although the random model performs well in estimating the total phenotypic variance and partitioning the additive and QTL heritabilities. When there is zero marker informativeness, the distribution method is weaker in detecting the presence of the QTL than it is without polygenic variance, yet it is still strong enough to identify its presence 26% of the time.

Discussion

Our results extend the general conclusions of Haley and Knott (1992) to the random model and show that the expectation method can capture most of the information available for QTL mapping. We find two important exceptions: (1) at the lower limit, when there is zero information in the marker loci, the distribution method can still successfully detect a QTL a high percentage of the time; and (2) the distribution method can be superior in a sib-pair–difference model. The distribution method is able to detect a QTL with even zero marker informativeness because, regardless of the marker alleles, in some populations the variance is partially explained purely by the decomposition of $f(y_i)$ into $f_0(y_i)$, $f_{1/2}(y_i)$, and $f_1(y_i)$. Under the sib-pair model the effect is weak, and it manifests itself only when the amount of information is very low. When there are no segregating markers, the expectation method is identical to the null hypothesis and is consequently powerless to detect the QTL.

We did not find a parameter space where there was nonzero marker informativeness yet still a significant benefit to using the distribution method in the sib-pair model. Despite this, it is feasible that there is a gradation of increasing information content such that the distribution method retains an advantage. As genetic maps become denser and the number of molecular markers increase, this point may be academic. Still, other problems, such as paternity assurance (i.e., cryptic half sibs in a full-sib model), may contribute degrees of uninformativeness that will warrant further directed analysis in light of the distribution-method approach.

Using the sib-pair model, the expectation method is relatively robust to the amount of marker information, although there is still a small problem in partitioning the heritabilities h_g^2 and h_a^2. This tends to be a general difficulty when either the sib-pair or the sib-pair–difference model is used (see, for example, Amos 1994; Amos et al. 1996). The problem is central to the robustness of the random model, since variance decomposition is analogous to the separation of mean effects in the fixed model. This is seen more clearly when information in trait values is reduced, such as when using the sib-pair–difference model. Here, the expectation method is considerably worse than the distribution method in partitioning the heritabilities. This means that there is an interaction effect in how information in the trait and marker loci is used. In the expectation/sib-pair–difference simulations, the sum $h_g^2 + h_a^2$ is no longer conserved, implying a confounding with the residual error term. We assume normally distributed error terms, and, although the maximum-likelihood method can be relatively robust to violations of this assumption, it is not entirely insensitive to it. Amos et al. (1996) recommend the quasi-likelihood technique when there is strong evidence that the assumption does not hold.

The sib-pair–difference simulations show no benefit in predicting the position of a QTL, nor in reducing

Table 5

The Sib-Pair Model with Nonzero Polygenic Heritability

	Expectation Model					Distribution Model				
	Position	CV	σ^2	h_g^2	h_a^2	Position	CV	σ^2	h_g^2	h_a^2
Standard	9.92 (5.78)	.58	24.94 (1.16)	.48 (.1545)	.26 (.1779)	10.03 (5.98)	.60	24.94 (1.16)	.46 (.1406)	.28 (.1684)
Minimum	18.63 (16.05)	.86	24.85 (1.66)	.51 (.2769)	.23 (.2692)	19.14 (16.71)	.87	24.85 (1.67)	.43 (.2381)	.31 (.2531)
Zero	. . . (. . .)	. . .	25.01 (1.12)	.55 (.1882)	.20 (.1774)	. . . (. . .)	. . .	25.02 (1.12)	.37 (.2551)	.38 (.2732)

NOTE.—The standard model is the same as in table 2, except that the non-QTL variance (12.5) is partitioned equally into polygenic and environmental components ($h_g^2 = .50$, $h_a^2 = .25$). The minimum model retains these heritabilities but sets all other parameters to their minimum information content (low marker informativeness, $N = 250$ families, 40-cM marker interval). The zero model is the same as the standard model, but with zero marker informativeness. The ellipses indicate that the zero model cannot provide information on the QTL position.

the variance in the estimate. One benefit of a sib-pair–difference procedure is its ability to compensate for common family effects. Yet, without data collected to isolate the effect by some differential contribution (e.g., by collecting data for both MZ and DZ twins), a common family effect will remain confounded with the other terms. This problem is made somewhat more difficult by the lack of analytical expressions for the distributions of allele-sharing information without resorting to pedigree analysis.

The problem of variance confounding is inherent in the requirements that model (1) makes on data acquisition. Although (1) is the stated model, the data are not appropriately organized to analyze it by, for example, standard ANOVA techniques. Model (1) implies that one can gather data separately by both QTL and polygenic effect (i.e., organize data appropriately for a two-way ANOVA without interaction) and then test the between- and within-group variances. Yet we assume the data consist solely of net effects y_{ij} organized by family and information on individuals' marker genotypes. As such, the data are amenable to, at most, the model

$$y_{ij} = \mu + f_i + \eta_{ij} \, , \qquad (5)$$

where f_i is the ith family effect and η_{ij} is the error term. Consequently, model (1) is overparameterized: there is not enough information in just the trait values themselves to partition the variance. Knott and Haley (1992) address this problem by constructing a maximum-likelihood equation that integrates (1) over a_i. Traditional genetic analyses circumvent this problem by implementing the probability of identity by descent as a covariate term that attempts to partition $2f_i$ into $g_i + a_i$. The factor 2 is incorporated because σ_f^2 (the between-group component) is equal to the total genetic component minus the covariance among group members (sibs). The partitioning of σ_g^2, σ_a^2, and σ_ε^2 is incorporated by the equality

$$\text{cov}(y_{i1},y_{i2}) = \sigma_{12,i} = \pi_i\sigma_g^2 + {}^1\!/_2\sigma_a^2 \, , \qquad (6)$$

where π_i is a random variable that takes values 0, $^1/_2$, and 1, depending on whether sibs y_{i1} and y_{i2} share zero, one, or two alleles at the putative QTL.

Table 6

Power to Detect a QTL when There Is Polygenic Variance

	Method	
Sib-Pair Model	Expectation	Distribution
Standard	.90	.90
Minimum	1.00	1.00
Zero	.01	.26

NOTE.—Each cell reports the proportion of runs that had at least one LR test statistic greater than the critical test statistic at the 95% level. Runs are as described in table 4.

As of yet, there is no formal theory that maps model (5) into model (1) via (6). That is, there are *methods* of using (1) and (6) to predicted QTL positions (for example, Goldgar 1990; Schork 1993), but there is not a formal QTL theory that generates a two-way ANOVA model from a one-way ANOVA model via correlations between effects (although, for a different approach to the problem, see Jansen [1992]). The consequence is that the practical ability to partition the variances σ_g^2 and σ_a^2 is dependent on the realized distribution of π_i. The relationship between these terms is incorporated in the correlation coefficient

$$\rho_i = \frac{\sigma_{12,i}}{\sigma^2} = \pi_i/h_g^2 + {}^1\!/_2 h_a^2 \, ,$$

where $\sigma^2 = \sigma_g^2 + \sigma_a^2 + \sigma_e^2$ includes the environmental variance σ_e^2, which includes, perhaps solely, the error variance σ_ε^2. When testing a specific putative position for a QTL, uncertainty in π_i for any given family reflects a lack of information. As the putative position departs from either marker, the lack of information is reflected by an increase in the variance in π_i within families and reduces the ability to partition σ_g^2, σ_a^2, and σ_e^2. The quantification of this is dependent on the magnitude of σ_g^2 relative to σ_a^2 and σ_e^2 and confirms our biological intuition that QTLs with low heritabilities are difficult to map. While low heritabilities have a mapping cost that is easy to understand in terms of a signal-to-noise motif, high heritabilities also impose a cost. On the assumption of constant population-wide heritabilities, the variance of ρ_i is equal to

$$\mathrm{var}(\rho_i) = (h_g^2)^2\mathrm{var}(\pi_i) ,$$

and, consequently, high heritabilities magnify the effect of within-family variance on ρ_i.

Across families, variance in π_i makes it more likely that regression, maximum likelihood, or other techniques will be able to map the QTL and separate the variance components. But even with adequate across-family variance, a successful partitioning of $(\sigma_g^2 + \sigma_a^2)$ and σ_e^2 does not guarantee an equally successful partitioning of σ_g^2 and σ_a^2. By extension, this applies in analogous ways to additional effects, such as common family, dominance, and epistatic effects.

This problem is well appreciated in the field. There are numerous statistical techniques to decompose variance components (Searle 1971; Cornell 1990; Searle et al. 1992), isolate a common family effect from other genetic variances (e.g., Elston 1988), and extend the GLM to genetic structural equations (Carey 1986). Most approaches to variance decomposition have developed theory in concert with experimental design so as to isolate variance contributions and then test their significance. Solutions of this sort are less applicable in the behavioral and social sciences and econometrics, and, accordingly, these fields have a long history of techniques for handling under- and overidentified structural equations (see, for example, the introduction by Goldberger [1973]). All the same, geneticists have been cautionary in adopting these methods, in part because of a number of disadvantages (Martin and Eaves 1977). In human behavioral genetics, this problem has long been recognized (Elston 1973), and various techniques have been used to estimate variance components, e.g., by contrasting the variance explained by differing groups of relatedness (Jinks and Fulker 1970; Fulker 1973; Loehlin 1978) or extending factor-analysis techniques to biometrical genetic models (Martin and Eaves 1977). To date, QTL analysis has not yet benefited from their implementation. While current methods can produce acceptable behavior under the random model, it seems that the real advances will come with a ground-up development of the random model that is sensitive to the data collection restrictions of human populations.

References

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535–543

Amos CI, Zhu DK, Boerwinkle E (1996) Assessing genetic linkage and association with robust components of variance approaches. Ann Hum Genet 60:143–160

Carey G (1986) A general multivariate approach to linear modeling in human genetics. Am J Hum Genet 39:775–786

Cornell JA (1990) Experiments with mixtures. John Wiley and Sons, New York

Elston RC (1973) Methodologies in human behavior genetics. Soc Biol 20:276–279

——— (1988) Human quantitative genetics. In: Weir BS, Eisen EJ, Goodman MM, Namkoong G (eds) Proceedings of the Second International Conference on Quantitative Genetics. Sinauer, Sunderland, MA, pp 281–282

Fulker DW (1973) A biometrical genetic approach to intelligence and schizophrenia. Soc Biol 20:266–275

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

Goldberger AS (1973) Structural equation models: an overview. In: Goldberger AS, Duncan OD (eds) Structural equation models in the social sciences. Seminar, New York, pp 1–18

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

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

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

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

Jansen RC (1992) A general mixture model for mapping quantitative trait loci by using molecular markers. Theor Appl Genet 85:252–260

Jinks JL, Fulker DW (1970) Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. Psychol Bull 73:311–349

Knott SA, Haley CS (1992) Maximum likelihood mapping of quantitative trait loci using full-sib families. Genetics 132: 1211–1222

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

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Loehlin JC (1978) Heredity-environment analyses of Jencks's IQ correlations. Behav Genet 8:415–436.

Martin NG, Eaves LJ (1977) The genetical analysis of covariance structure. Heredity 38:79–95

Martínez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85:480–488

Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:308–313.

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete RFLP linkage map. Nature 335:721–726

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

Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris.* Genetics 8: 552–560

Schork NJ (1993) Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. Am J Hum Genet 53:1306–1319

Searle SR (1971) Linear models. John Wiley and Sons, New York

Searle SR, Casella G, McCulloch CE (1992) Variance components. John Wiley and Sons, New York

Xu S (1995) A comment on the simple regression method for interval mapping. Genetics 141:1657–1659

Xu S, Atchley WR (1995) A random model approach to interval mapping of quantitative trait loci. Genetics 141:1189–1197