A Random Model Approach to Interval Mapping of Quantitative Trait Loci

Shizhong Xu* and William R. Atchley†

**Department of Botany and Plant Sciences, University of California, Riverside, California 92521-0124 and †Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695-7614*

Manuscript received April 4, 1995
Accepted for publication August 2, 1995

ABSTRACT

Mapping quantitative trait loci in outbred populations is important because many populations of organisms are noninbred. Unfortunately, information about the genetic architecture of the trait may not be available in outbred populations. Thus, the allelic effects of genes can not be estimated with ease. In addition, under linkage equilibrium, marker genotypes provide no information about the genotype of a QTL (our terminology for a single quantitative trait locus is QTL while multiple loci are referred to as QTLs). To circumvent this problem, an interval mapping procedure based on a random model approach is described. Under a random model, instead of estimating the effects, segregating variances of QTLs are estimated by a maximum likelihood method. Estimation of the variance component of a QTL depends on the proportion of genes identical-by-descent (IBD) shared by relatives at the locus, which is predicted by the IBD of two markers flanking the QTL. The marker IBD shared by two relatives are inferred from the observed marker genotypes. The procedure offers an advantage over the regression interval mapping in terms of high power and small estimation errors and provides flexibility for large sibships, irregular pedigree relationships and incorporation of common environmental and fixed effects.

THERE are two primary types of data used for mapping a quantitative trait locus (QTL): data derived from inbred lines that include back cross, F_2 or more derived populations and field collected data such as those sampled from human populations. With data from line crosses, the parental genotypes, the linkage phases of loci and the number of alleles of the putative QTL are known precisely. In addition, such data from designed experiments can be considered as being from one large family because all individuals share the same parental genotypes. As a result, the effect of QTL substitution and dominance are directly estimated (LANDER and BOSTEIN 1989; HALEY and KNOTT 1992; ZENG 1994). The linear model describing such data is a fixed model. With field data, however, the parental genotypes may not be known. In addition, there will be many different families and the probability of a QTL genotype conditional on a marker genotype will differ from family to family. The linkage phases of parents are usually not known. Although they can be inferred from genotypes of family members, the family size is usually not large enough to allow accurate estimation. As a result, one must try all possible linkage phases and choose the one with the greatest probability (KNOTT and HALEY 1992). Furthermore, the number of alleles of the putative QTL is never known if the origin of the population is unclear. Corresponding to the complicated situation with field data, robust methods, based on a random model approach, were developed (HASEMAN and ELSTON 1972; AMOS and ELSTON 1989; AMOS *et al.* 1989) where knowledge of the actual genetic model of the QTL is not absolutely required.

Mapping quantitative trait loci with data derived from crosses between two outbred populations is occasionally possible, but these analyses are more difficult than those with other types of data. A population derived from such crosses is in linkage disequilibrium, which violates the assumption required by the random model approach. The uncertainty in the number of alleles and linkage phases of linked loci leads to serious difficulties when a fixed model approach is used. Assuming alternative fixation for QTLs in two diverged outbred lines, the least squares approach (fixed model) of HALEY *et al.* (1994) for gene mapping can be used. Mapping QTLs in data from crosses between outbred lines has been carried out in pigs by ANDERSSON *et al.* (1994).

The random model approach to gene mapping in outbred populations (also referred to as random populations because linkage equilibrium is assumed) are based on differentiated proportions of genes identical-by-descent (IBD) shared by two relatives. The assumption is that the greater the shared proportion of IBD, the more similar the phenotypes of the two relatives. Indeed, HASEMAN and ELSTON (1972) proved the negative linear relationship between the squared phenotypic difference and the IBD proportion. This permits use of a simple regression analysis to detect linkage. HASEMAN and ELSTON (1972) regressed the squared phenotypic difference between sibs on the shared IBD and the regression coefficient is negatively proportional to the variance explained by the QTL.

Corresponding address: Shizhong Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521-0124. E-mail: xu@genetics.ucr.edu

However, the genetic variance and linkage parameter are confounded in the Haseman-Elston sib-pair method. A sib-pair interval mapping procedure was recently developed by FULKER and CARDON (1994) using two flanking markers simultaneously to separate these two terms and to locate the quantitative trait locus at a specific position on a chromosome. Although statistical power has been improved, it is still a least squares based method and therefore does not optimally utilize information from the data. GOLDGAR (1990) developed a multipoint IBD method to estimate the total amount of genetic material shared by relatives in a given chromosomal region and eventually used a maximum likelihood (ML) approach to estimate the genetic variance explained by that particular region. GOLDGAR's ML is a general method that can be used for any number of sibs or irregular pedigree relationships. This method was extended by SCHORK (1993) to simultaneously estimate variances of several chromosomal regions and the common environment shared by family members. The ML method takes advantage of the distributional property of the data and therefore is more efficient than the Haseman-Elston test (GOLDGAR 1990; GOLDGAR and ONIKI 1992). Both regression (HASEMAN and ELSTON 1972; FULKER and CARDON 1994) and ML analyses (GOLDGAR 1990) estimate the variance associated with a QTL (or a chromosomal region). Therefore, the models used in these two methods are random models. SCHORK (1993) and VAN ARENDONK *et al.* (1994) also considered the same problem from a mixed model perspective.

Although GOLDGAR (1990) used two flanking markers to define a chromosomal segment and a ML method to estimate the genetic variance, his method was not designed for interval mapping. Rather, it is intended to test whether at least one QTL is located somewhere in the region. It has been suggested that GOLDGAR's ML method can be used as a first step in mapping QTLs (GOLDGAR and ONIKI 1992). When significant variation due to a particular region is detected by GOLDGAR's method, other approaches could be used to characterize the specific position of the QTL. If there are large numbers of markers covering the whole genome, the interval flanked by two adjacent markers is expected to be small, *e.g.*, 10 or 15 cM. With a sparse distribution of QTL positions, *i.e.*, there are only a few QTLs randomly distributed along a chromosome, the chance of two QTLs occurring in the same interval may be negligible. In this case, it is feasible to use GOLDGAR's ML method for interval mapping. Interval mapping using the ML method will be more efficient than the sib-pair regression method if the distributional properties of the data are known.

The aim of this study is to develop a general QTL mapping procedure using a random model approach (estimating variance) for outbred (random) populations. We extend GOLDGAR's (1990) ML method to interval mapping and use Monte Carlo simulation to compare the efficiency and statistical power of this new interval mapping procedure with the existing regression mapping (FULKER and CARDON 1994).

THEORY AND METHODS

Herein, we introduce two models: a single QTL model where one QTL is assumed on a tested chromosome and a multiple QTL model where more than one QTLs exist on the tested chromosome.

Single QTL model: The random model is defined by GOLDGAR (1990) as

$$y_{ij} = \mu + g_{ij} + a_{ij} + e_{ij} \quad (1)$$

where y_{ij} is the phenotypic value of the jth member in the ith family, μ is the population mean, $g_{ij} \sim N(0, \sigma_g^2)$ is the additive genetic effect (random) of the QTL to be tested on a chromosome, $a_{ij} \sim N(0, \sigma_a^2)$ is the polygenic additive effect (random) and $e_{ij} \sim N(0, \sigma_e^2)$ is the environmental deviation. The polygenic term is the summation of effects of all trait loci located on other chromosomes (excluding the putative QTL). Dominance effects are ignored here for simplicity. Note that other random effects, such as common environmental effect, can be easily incorporated into the model, but we have chosen this simple model solely to demonstrate the maximum likelihood interval mapping procedure.

The random model is generalized for any pedigree relationship within a family, but to diminish the confusion caused by complicated notation for arbitrary pedigree relationships, only full-sibs are considered in this presentation. Under the random model, $E(y_{ij}) = \mu$. Assuming linkage equilibrium, the variance of y_{ij} is

$$\mathrm{Var}(y_{ij}) = \sigma^2 = \sigma_a^2 + \sigma_g^2 + \sigma_e^2 \quad (2)$$

The coviariance between two noninbred sibs is

$$\mathrm{Cov}(y_{ij}, y_{ij'}) = \pi_{iq}\sigma_g^2 + \tfrac{1}{2}\sigma_a^2 \quad (3)$$

where π_{iq} is the proportion of alleles IBD shared by family member j and j' at the putative QTL. The coefficient of the polygenic variance in (3) is $\frac{1}{2}$ because, by expectation, two noninbred sibs share $\frac{1}{2}$ genes IBD. The IBD of the QTL, π_{iq}, will be different from one sib-pair to another. This is fundamentally different from the polygenic treatment of a quantitative trait where the IBD value always takes $\frac{1}{2}$. If we do not observe or do not have any information about the genotypes of the two sibs for the trait locus, it is natural to replace π_{iq} by its expectation, *i.e.*, $\frac{1}{2}$. However, the actual π_{iq} is a variable that ranges from 0 to 1.

Let us consider the genotypic configurations of progenies from the mating type, $A_1A_2 \times A_3A_4$. There are four possible types of progeny, each with an equal frequency. The four possible genotypes are $\frac{1}{4}A_1A_3$, $\frac{1}{4}A_1A_4$, $\frac{1}{4}A_2A_3$ and $\frac{1}{4}A_2A_4$. If two sibs are sampled from this family, ignoring the sampling order, there are 10

possible sib pairs. Suppose we observe two sibs with A_1A_3-A_1A_3. We know they have received exactly the same alleles from their parents and the IBD is 1. The two sibs behave just like identical twins for this locus. If we observe A_1A_3-A_2A_4, then we know they do not share any IBD alleles, thus behave like two unrelated individuals. This means that if we happen to know the genotypes of two sibs at a particular locus, the covariance between sibs at this locus conditional on the genotypes may be different from what is expected. For example, $\mathrm{Cov}(y_{ij},y_{ij'}) = 1\ \sigma_g^2 + \frac{1}{2}\sigma_a^2$ for a pair of sibs with genotypic configuration of A_1A_3-A_1A_3 at the QTL and $\mathrm{Cov}(y_{ij},y_{ij'}) = 0\ \sigma_g^2 + \frac{1}{2}\sigma_a^2$ with A_1A_3-A_2A_4. It is incorrect to use $\mathrm{Cov}(y_{ij},y_{ij'}) = \frac{1}{2}\sigma_g^2 + \frac{1}{2}\sigma_a^2$ if we already know that j and j' share no IBD at the QTL.

In practice, genotypes of QTL are not observable. However, we can observe the genotypes of markers linked to the QTLs. HASEMAN and ELSTON (1972) developed the joint probability for two linked loci and showed that the expected IBD of one locus is a linear function of the IBD of another locus. FULKER and CARDON (1994) proposed IBD of two flanking markers to calculate the conditional mean of π_{iq}. The conditional mean of π_{iq} is also a linear function of πs of the two flanking markers (FULKER and CARDON 1994). Let θ_{12} be the recombination fraction between the flanking markers while $\theta_1 q$ and θq_2 are recombination fractions between the trait locus and marker 1 and marker 2, respectively. Then

$$\hat{\pi}_{iq} = \mathrm{E}(\pi_{iq}|\pi_{i1}\ \pi_{i2}) = \alpha + \beta_1\pi_{i1} + \beta_2\pi_{i2} \qquad (4)$$

where π_{i1} and π_{i2} are the IBD values of the two flanking markers. FULKER and CARDON (1994) showed

$$\beta_1 = [(1 - 2\theta_{1q})^2 - (1 - 2\theta_{q2})^2(1 - 2\theta_{12})^2]/[(1 - (1 - 2\theta_{12})^4];$$

$$\beta_2 = [(1 - 2\theta_{q2})^2 - (1 - 2\theta_{1q})^2(1 - 2\theta_{12})^2]/[(1 - (1 - 2\theta_{12})^4];$$

$$\alpha = (1 - \beta_1 - \beta_2)/2. \qquad (5)$$

The term π_{iq} in the covariance given in Equation 3 is substituted by this conditional mean ($\hat{\pi}_{iq}$) when estimation of variances is performed.

With two sibs in the ith family, for example, the covariance matrix is

$$\mathbf{V}_i = \mathrm{Var}\begin{bmatrix} y_{ij} \\ y_{ij'} \end{bmatrix} = \sigma^2\begin{bmatrix} 1 & r_i \\ r_i & 1 \end{bmatrix} \qquad (6)$$

where

$$r_i = \hat{\pi}_{iq}h_g^2 + \tfrac{1}{2}h_a^2$$

Here we define $h_g^2 = \sigma_g^2/\sigma^2$ as the heritability of the putative QTL and $h_a^2 = \sigma_a^2/\sigma^2$ the heritability of the polygenic effect. Let us first define $\mathbf{C}_i$ as

$$\mathbf{C}_i = \begin{bmatrix} 1 & r_i \\ r_i & 1 \end{bmatrix}$$

If there are k sibs in each family, C_i is simply a $k \times k$ matrix. If normal distribution of the data ($\mathbf{y}$) is assumed, we have the following joint density function of observing a particular vector of data,

$$f(\mathbf{y}_i) = \frac{1}{(2\pi\sigma^2)^{k/2}|\mathbf{C}_i|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_i - \mathbf{1}\mu)^T\mathbf{C}_i^{-1}(\mathbf{y}_i - \mathbf{1}\mu)\right\} \qquad (7)$$

where $\mathbf{y}_i = [y_{i1}, \ldots, y_{ik}]^T$ is a $k \times 1$ vector of phenotypes of the ith sibship and k is the family size, and $\mathbf{1}$ is a $k \times 1$ vector with all entries equal to 1. In fact, normal distribution of the QTL effect is not absolutely required as long as the QTL variance is small compared with the sum of the polygenic and environmental variances, but normality of a_{ij} and e_{ij} is required to make $\mathbf{y}_i$ normal. The overall log likelihood for n independent families is

$$L = \sum_{i=1}^{n} \log[f(\mathbf{y}_i)] \qquad (8)$$

This likelihood function relates to the position of the trait locus flanked by the two markers through r_i. The unknown parameters are μ, σ^2, h_g^2, h_a^2 and θ_{1q}. Common practice of interval mapping is not maximizing L with respect to all the parameters but first treating θ_{1q} as a known constant, then varying θ_{1q} throughout the entire interval, and eventually every interval in the whole genome. The maximum likelihood estimate of the QTL position takes the value of θ_{1q} in the appropriate interval that maximizes L throughout the entire chromosome.

At any particular position in an interval, the algorithm employed here does the following: first, for given prior h_g^2 and h_a^2, the maximum likelihood estimates (MLEs) of μ and σ^2 can be expressed as functions of h_g^2 and h_a^2, *i.e.*,

$$\hat{\mu} = u(h_g^2, h_a^2) = \left[\sum_{i=1}^{n} \mathbf{1}'\mathbf{C}_i^{-1}\mathbf{1}\right]^{-1}\left[\sum_{i=1}^{n} \mathbf{1}'\mathbf{C}_i^{-1}\mathbf{y}_i\right] \qquad (9)$$

and

$$\hat{\sigma}^2 = v(h_g^2, h_a^2) = \frac{1}{nk}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{1}\hat{\mu})^T\mathbf{C}_i^{-1}(\mathbf{y}_i - \mathbf{1}\hat{\mu}) \qquad (10)$$

If family size varies, nk should be replaced by $\Sigma_{i=1,\ n}k_i$. These two equations are obtained by setting the partial derivatives of L with respect to μ and σ^2 equal to zeros, respectively.

Second, substitute μ and σ^2 in the original likelihood function (Equation 8) by their MLEs, so that the log likelihood is expressed as a function of h_g^2 and h_a^2, as given below

$$L = -\frac{1}{2} nk \log(\hat{\sigma}^2) - \frac{1}{2} \sum_{i=1}^{n} \log(|\mathbf{C}_i|) \quad (11)$$

Note that after the substitution, μ and σ^2 are absorbed in the likelihood function. We do not treat μ and σ^2 as fixed, rather, we replace them by (9) and (10), respectively. Here we directly maximize (11) with respect to h_g^2 and h_a^2 via the simplex algorithm (NELDER and MEAD 1965). It should be noted that substitution of μ and σ^2 by u and v implies that when h_g^2 and h_a^2 are updated, the two functions, $\hat{\mu} = u(h_g^2, h_a^2)$ and $\hat{\sigma}^2 = v(h_g^2, h_a^2)$, are also updated.

In fact, we can directly use the simplex algorithm to maximize the likelihood function given in (8) with respect to all four parameters (μ, σ^2, h_g^2 and h_a^2) simultaneously, but it is computationally inferior because the dimensionality of unknowns will be four instead of two.

Symbolically, let us express Equations 8 and 11 by

$$L_1 = \text{function}(\mu, \sigma^2, \mathrm{h}_g^2, \mathrm{h}_a^2) \quad (8a)$$

and

$$L_2 = \text{function}[u(\mathrm{h}_g^2, h_a^2), v(\mathrm{h}_g^2, \mathrm{h}_a^2), \mathrm{h}_g^2, \mathrm{h}_a^2], \quad (11a)$$

respectively. The algorithm presented here is to maximize L_2 with respect to two parameters (h_g^2 and h_a^2), which is equivalent to maximizing L_1 with respect to four parameters (μ, σ^2, h_g^2 and h_a^2). A similar dimension-reduction technique has been used by GRASER *et al.* (1987).

The null hypothesis is H_0:$h_g^2 = 0$, *i.e.*, there is no QTL segregating in the tested interval. The ML under the null hypothesis is denoted by L_0. The likelihood ratio (LR) test statistic is

$$\text{LR} = -2(L_0 - L_1) \quad (12)$$

which follows a chi-square distribution with $2 \geq$ d.f. $>$ 1 under H_0. One degree of freedom is due to fitting h_g^2 and the remaining degree of freedom is for fitting the QTL position (HALEY and KNOTT 1992). For a given interval, the remaining degree of freedom depends on the size of the interval (θ_{12}) and it is less than one because we only search the QTL within the tested interval, rather than the entire genome. If the null hypothesis is no QTL in the whole genome (not just one interval) covered by the markers, then df $\approx$ 2 under the null hypothesis. With many markers on a long chromosome, the number of degrees could be greater than two (ZENG 1994) for a chromosomal wise test.

Multiple QTL model: The above model assumes there is only one QTL in the linkage group where the tested interval is located. If there are multiple QTLs in the same linkage group, the estimation tends to be biased because of interference caused by QTLs located on the same chromosome but outside the tested interval (HALEY and KNOTT 1992; MARTINEZ and CURNOW 1992; JANSEN 1993, 1994; ZENG 1993, 1994). The multiple QTL model is described by

$$y_{ij} = \mu + a_{ij} + \sum_{k=1}^{S} u_{ij}^k + g_{ij} + \sum_{r=1}^{W} u_{ij}^r + e_{ij} \quad (13)$$

where u_{ij}^k and u_{ij}^r are the kth QTL effect on the left side and the rth QTL effect on the right side of the putative QTL, S and W are the numbers of trait loci in the left and right sides of the current QTL. Under the assumption of linkage equilibrium, the variance of y_{ij} is

$$\text{Var}(y_{ij}) = \sigma^2 = \sigma_a^2 + \sum_{k=1}^{S} \sigma_k^2 + \sigma_g^2 + \sum_{r=1}^{W} \sigma_r^2 + \sigma_e^2 \quad (14)$$

and the covariance between noninbred sibs is

$$\text{Cov}(y_{ij}, y_{ij'}) = \tfrac{1}{2} \sigma_a^2 + \sum_{k=1}^{S} \pi_{ik}\sigma_k^2 + \pi_{iq}\sigma_g^2 + \sum_{r=1}^{W} \pi_{ir}\sigma_r^2 \quad (15)$$

Note that these πs are the IBD proportions of the QTLs, and they are not observable.

SCHORK (1993) proposed a similar model and suggested to include more chromosomal regions into the conditional covariance for purpose other than interval mapping. Theoretically, we can define the conditional covariance given the πs of all markers in the linkage group and include all variance components explained by each marker into the likelihood function to control other QTL effects. However, technically this is not feasible because of so many variance components and also it is not necessary. The IBD variable has the same property as the indicator variables (ZENG 1993) in that conditional on the π of a marker locus, the π of a QTL on one side of the marker is not correlated to that of a QTL on the other side. This will become evident when we reexamine Equation 5 by treating the QTL as a neutral marker. Let us define the two flanking markers and the QTL by i, j and k, respectively. Equation 5 is then rewritten as

$$\beta_1 = [(1 - 2\theta_{ik})^2 - (1 - 2\theta_{kj})^2(1 - 2\theta_{ij})^2]/[(1 - (1 - 2\theta_{ij})^4];$$

$$\beta_2 = [(1 - 2\theta_{kj})^2 - (1 - 2\theta_{ik})^2(1 - 2\theta_{ij})^2]/[(1 - (1 - 2\theta_{ij})^4]$$

This equation holds regardless of the order the three loci are arranged on the chromosome. If the order is $i < k < j$, this equation is required to predict π_k. However, if the order is $k < i < j$, then $(1 - 2\theta_{jk})^2 = (1 - 2\theta_{ik})^2(1 - 2\theta_{ij})^2$, which leads to $\beta_1 = (1 - 2\theta_{ik})^2$ and $\beta_2 = 0$. On the other hand, if $i < j < k$, then $(1 - 2\theta_{ik})^2 = (1 - 2\theta_{jk})^2(1 - 2\theta_{ij})^2$ leading to $\beta_1 = 0$ and $\beta_2 = (1 - 2\theta_{jk})^2$. This means that conditional on IBD of a marker in the middle, the IBD of a locus on one side provides no information about the IBD of a locus on the other side. Furthermore, conditional on two flanking loci, the IBD of an intermediate locus provides no information about the IBD values of loci outside the flanking region. These properties have been found in

ZENG (1994) for the indicator variables. The only difference is that the indicator variables are observed without error while the IBD proportions are estimated with some uncertainty if the mating type is not fully informative.

These properties are the bases of the composite interval mapping of ZENG (1993, 1994) and JANSEN (1993, 1994) for line cross populations, which will be directly adopted here for the random model approach. Theoretically, one marker is enough to block correlation between a locus on the left and a locus on the right. Therefore, we only need two additional markers flanking the current interval to block interference caused by outside QTLs. Here we still use markers 1 and 2 to denote the two flanking markers that define the tested interval, but use L and R to denote the next-to-flanking markers. Now the four markers have the order: $L - 1 - 2 - R$. The tested QTL is between markers 1 and 2. Let π_{iL} and π_{iR} be the IBD values of the left (L) and right (R) additional markers, respectively, and θ_{kL} and θ_{rR} be the recombination fractions of the kth QTL with locus L and the rth QTL with locus R. Given the four markers the conditional covariance between sibs is

$$\mathrm{Cov}(y_{i1}, y_{i2} | \pi_{iL}\hat{\pi}_{iq}\pi_{iR}) = \sigma^2[\pi_{iL}H_L^2 + \hat{\pi}_{iq}h_g^2 + \pi_{iR}H_R^2 + \tfrac{1}{2}h_a^2] \quad (16)$$

where $H_L^2 = \Sigma_{k=1}^{S}$, $(1 - 2\theta_{kL}^2)^2 h_k^2$, $H_R^2 = \Sigma_{r=1}^{W}$, $(1 - 2\theta_{rR}^2)^2 h_r^2$, $h_k^2 = \sigma_k^2/\sigma^2$ and $h_r^2 = \sigma_r^2/\sigma^2$. At a particular position, the parameters in the likelihood function are μ, σ^2, H_L^2, h_g^2, H_R^2 and h_a^2, but only h_g^2 is tested.

The computing algorithm follows that for single QTL model, but now there are more unknown variance components. The solutions of the unknowns must be searched within the appropriate solution space, otherwise the covariance matrix is not assured to be positive definite. A method of reparameterization described in the APPENDIX is used to obtain a positive definite covariance matrix and thus guarantee convergence to a solution.

There will be some interference if QTLs exist between markers L and 1 or R and 2. However, under the assumption of a dense marker map and a few QTLs randomly distributed along a chromosome, there may be small chance of two QTLs existing in two adjacent intervals. Theoretically, the tests of different intervals are independent (ZENG 1993), though independence may not be guaranteed with small population sizes.

The multiple QTL method has both advantages and disadvantages when compared with the single QTL one. The advantage is when the other intervals do contain QTLs whose effects can then be absorbed by the next-to-flanking markers. The disadvantage is when there is no QTL in the other intervals, in which case we loose power because the next-to-flanking markers will pick up random noise instead of QTL effects. In other words, by using the multiple QTL model we may gain accuracy at the price of loosing power.

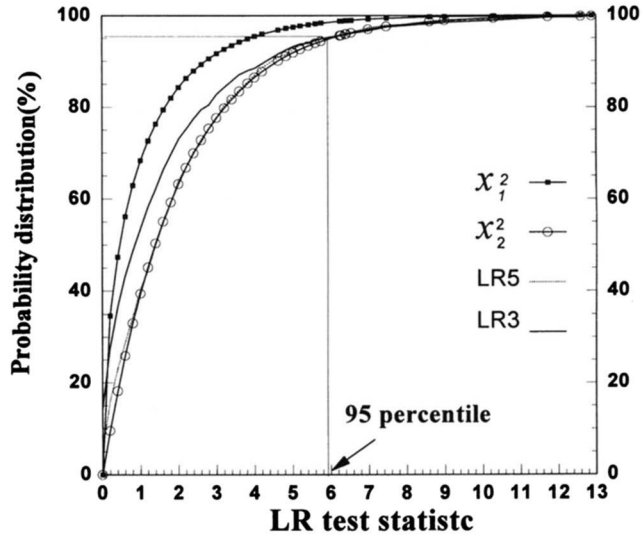

FIGURE 1.—Empirical cumulative distributions of the likelihood ratio (LR) test statistic. The curve is compared with the distributions of $\chi^2_{df=1}$ and $\chi^2_{df=2}$.

SIMULATION STUDIES

Test statistic under the null hypothesis: To further investigate the behavior and the threshold value of the test statistic of the ML method, we simulated data under the null hypothesis of no QTL on the tested chromosome. Under the null hypothesis, we simulated the polygenic effect with a heritability of 0.5. Six codominant markers each with six equally frequent alleles were simulated. The six markers were linked 20 cM apart and covered a linkage group of 100 cM length. The IBD values of markers were calculated using the method of HASEMAN and ELSTON (1972). Five hundred independent full-sib families, each with two sibs, were simulated in each experiment. The simulation experiment was repeated 1000 times. In each experiment, the maximum LR of the single QTL model was recorded throughout the entire chromosomal segment. The empirical distribution of the LR test statistic over 1000 replicates was examined and shown in Figure 1. The 95th percentile of the empirical distribution was 5.85. This figure also gives two chi-square distributions with 1 and 2 d.f., respectively. The empirical LR test statistic has a distribution almost indistinguishable from the $\chi^2_{df=2}$ distribution. Therefore, the critical value of 5.99 ($= \chi^2_{0.05(2)}$) was used to determine the level of significance in subsequent simulation studies.

To compare our method with FULKER and CARDON's (1994) regression approach, we also calculated the empirical distribution of the t statistic for the same set of simulated data. We found that the 95th percentile of the t statistic (one-tail test) was -2.448 for a chromosomewise test. Note that the critical value reported by

FULKER and CARDON (1994) was −1.856 for an intervalwise test. Because we were interested in chromosomewise test, the critical value −2.448 was used in this study. Note that the t statistic is approximately the square root of $\chi^2_{df=2}$ under the null hypothesis. This becomes obvious by looking at $(-2.448)^2 = 5.993$, which is virtually identical to the critical value of $\chi^2_{df=2}$ distribution. Therefore, instead of t, the t^2 test statistic was used for the sib-pair regression analysis. Before converting t into t^2 statistic, we replaced any positive value of t by 0 so that the t^2 test is still a one-tail test. In subsequent analyses, the threshold value of 5.99 was used for both the LR and t^2 tests.

Experimental design: Extensive simulation was done on the single QTL model. The genetic model is as follows: there was one QTL with either two or six allelic states of equal frequency. The allelic effects were set such that the additive variance explained by the QTL was 12.5 squared units. Dominance effect was assumed to be absent. No polygenic effect was simulated, but the polygenic term was fitted in the model when the ML method was used. Heritability was set at 0.25 and 0.50 by adjusting the amount of random environmental deviation assigned to the phenotype. Six codominant markers each with six alleles of equal frequency were simulated. The six markers were linked 20 cM apart and covered a linkage group of 100 cM length. The simulated QTL was located in the middle of the chromosome segment, *i.e.,* 50 cM. Two sibs in each family were simulated, and the number of families (sample size) varied at 250, 500 and 1000. In each set of parameter combination, the simulation was repeated 100 times.

Results: Result from a single replicate of simulation was depicted in Figure 2, where the parameters were: heritability = 0.50, number of alleles = 2, sample size = 500. The test statistic of the ML method was compared with that of the regression (RG) method, showing that the LR profile had a higher peak than the t^2 profile.

Mean estimates and standard deviations (over 100 replicates) of the QTL location (cMA), variance explained by the QTL (σ^2_g) and the polygenic variance (σ^2_a) are summarized in Table 1, which demonstrates: (1) the number of alleles in the QTL had little influence on estimation of both methods; (2) standard deviations of parameter estimates in ML were smaller than those in the regression method, *i.e.,* ML had smaller estimation errors than RG; and (3) QTL variance was overestimated by the regression method, especially when heritability and sample size were small.

Clearly, the maximum likelihood method performs much better that the regression method.

It is well known that the regression method generates unbiased estimates of regression coefficients. At first glance it might seem that estimation of the QTL variance should be unbiased with the regression method because it is estimated by a term proportional to the regression coefficient. However, the property of unbiasedness only holds when the QTL position is fixed. If the QTL location varies, then the variance associated with the QTL tends to be overestimated because the method always chooses a location with the highest t^2 that is proportional to the squared QTL variance.

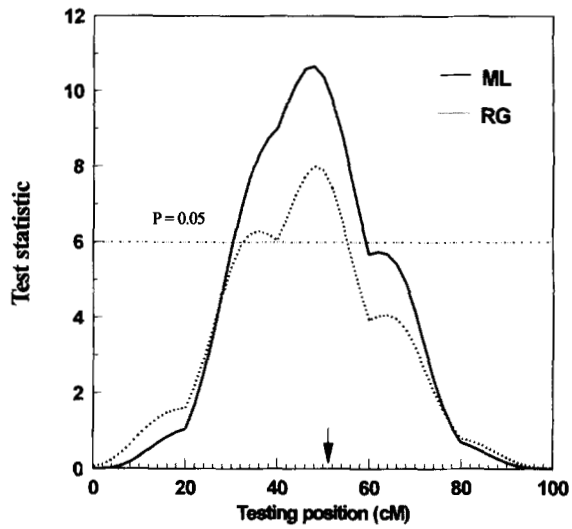

FIGURE 2.—Comparison of the LR profile of maximum likelihood (ML, ——) with the t^2 statistic profile of regression (RG, · · · · ·) drawn from a single simulation experiment of 500 full-sib families each with two members. The QTL is located at 50 cM, has two alleles and explains 50% of the total phenotypic variation.

The average test statistics and power estimates (at $\alpha = 0.05$) over 100 replicates are summarized in Table 2. The test statistics were likelihood ratio (LR) for the ML method and t^2 for RG method, both having a critical value of 5.99 under the null hypothesis. The ML method in general has a higher statistical power than the regression method.

Multiple QTL model: Because intensive computation is involved for the multiple QTL model, we only simulated one sample of 1000 sib-pairs for the multiple QTL model to demonstrate the behavior of the LR profile. More extensive simulation is left for a later paper. In this paper, we simulated two QTLs located in the 15 and 45 cM positions on a linkage group of size 50 cM. The two QTLs have an equal effect on the trait with a total heritability of 0.5. The result from the multiple QTL model was compared with that of the single QTL model. These LR profiles are shown in Figure 3. The multiple QTL model clearly indicates two QTLs at the correct positions while the single QTL model shows two major peaks, but each with two subpeaks. The two subpeaks in the middle are the "ghost images" (HALEY and KNOTT 1992) of the QTLs in the opposite sides. With the multiple QTL model approach, these ghost images are removed and the positions of the two QTLs are precisely mapped on the chromosome. The behavior of the LR test statistic under this model is unknown.

TABLE 1

Comparison of maximum likelihood and regression analyses of simulated data

Sample size	Heritability	Comparison	Two alleles			Six alleles		
			cM_A	$\hat{\sigma}_g^2$	$\hat{\sigma}_a^2$	cM_A	$\hat{\sigma}_g^2$	$\hat{\sigma}_a^2$
250	0.25	ML	49.66 (26.61)	12.64 (5.39)	1.61 (4.16)	49.90 (27.02)	12.54 (5.19)	1.63 (3.88)
		RG	49.84 (28.48)	21.36 (11.09)	—	52.00 (28.16)	21.56 (11.71)	—
	0.50	ML	49.44 (19.15)	11.56 (2.97)	1.99 (3.09)	50.98 (19.08)	11.69 (2.91)	1.36 (2.74)
		RG	49.10 (19.58)	14.35 (5.65)	—	52.74 (19.62)	14.63 (5.66)	—
500	0.25	ML	46.12 (24.53)	12.05 (4.25)	1.53 (3.33)	48.46 (23.43)	11.78 (4.34)	1.54 (3.43)
		RG	45.02 (26.85)	18.04 (8.41)	—	46.18 (26.43)	17.94 (8.52)	—
	0.50	ML	49.32 (12.64)	11.67 (2.45)	1.20 (2.23)	51.94 (14.42)	11.55 (2.49)	1.12 (2.17)
		RG	50.74 (14.04)	13.75 (4.33)	—	49.18 (15.30)	13.69 (4.43)	—
1000	0.25	ML	46.62 (23.77)	11.34 (3.57)	1.74 (3.18)	53.28 (20.28)	12.26 (3.37)	1.27 (2.94)
		RG	45.92 (24.04)	15.48 (7.29)	—	51.80 (23.19)	17.91 (6.94)	—
	0.50	ML	49.26 (6.36)	11.83 (1.85)	0.68 (1.44)	48.60 (8.68)	11.59 (1.76)	0.94 (1.68)
		RG	49.42 (8.15)	13.60 (3.55)	—	46.86 (11.80)	13.49 (3.33)	—

ML, maximum likelihood; RG, regression; cM_A, QTL location; $\hat{\sigma}_g^2$, variance explained by the QTL; $\hat{\sigma}_a^2$, polygenic variance. The standard deviations of the estimates over 100 replicates are given in parentheses. The true location (cM_A) and QTL variance (σ_g^2) are 50 cM and 12.5 squared units, respectively, while the true polygenic variance is zero.

If we had used the critical value from the single QTL model, the power would have been reduced with the multiple QTL model. By using the multiple QTL model, we may gain accuracy at the cost of power. Similar results are present in the composite interval mapping procedures of ZENG (1994).

DISCUSSION

Recent studies that compare ML with regression in line crosses find virtually no difference between the two methods (HALEY and KNOTT 1992; MARTINEZ and CURNOW 1992). At first glance it might seem that our ML should not be significantly different from FULKER and CARDON's (1994) regression. However, the ML and regression compared by HALEY and KNOTT (1992) and others are based on a fixed model approach that directly estimates the QTL effect. The ML and regression compared in this study deal with a random model that estimates the QTL variance. Under some circumstances, *e.g.*, the QTL being at one flanking marker, estimates of the QTL effect by the regression and ML are equivalent under the fixed model. If the QTL is not at a flanking marker, the difference is expected to be negligible if the QTL effect is small relative to the residual standard deviation. Therefore, regression mapping is an approximation of ML mapping under a fixed model. In contrast to the fixed model, random model regression mapping is not an approximation of ML under any circumstances. The two methods may generate similar results if sample sizes are large, but there is no mathematical basis for any equivalence. The higher power of ML compared with regression is probably due to the fact that ML uses the phenotypic values as the

TABLE 2

Comparison of statistical powers between maximum likelihood and regression analyses of simulated data

Sample size	Heritability	Comparison	Two alleles		Six alleles	
			Test statistic	Power (%)	Test statistic	Power (%)
250	0.25	ML	2.70 (1.95)	08	2.65 (1.94)	06
		RG	2.55 (2.04)	10	2.62 (2.16)	09
	0.50	ML	6.63 (3.74)	52	6.59 (3.68)	51
		RG	5.35 (3.36)	35	5.37 (3.51)	31
500	0.25	ML	3.94 (2.56)	17	3.82 (2.69)	19
		RG	3.55 (2.66)	14	3.55 (2.63)	15
	0.50	ML	11.46 (5.23)	83	11.22 (5.43)	83
		RG	8.88 (4.92)	71	8.70 (4.90)	71
1000	0.25	ML	6.05 (4.17)	44	6.92 (3.82)	54
		RG	4.93 (4.24)	27	6.04 (4.19)	42
	0.50	ML	20.61 (7.47)	96	19.94 (8.03)	100
		RG	16.36 (7.62)	91	15.18 (6.57)	96

Values are mean estimates of the test statistic and power at an error rate of 0.05 (with standard deviation of the estimate over 100 replicates in parentheses).

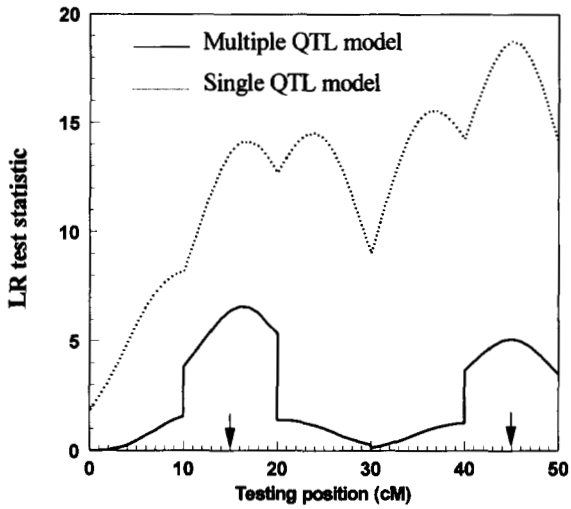

FIGURE 3.—Comparison of the LR profile of the multiple QTL model (——) with that of the single QTL one (· · · · ·). Results were drawn from a single simulation experiment of 1000 full-sib families each with two members. The two QTLs, accounting for 50% of the total phenotypic variation, are located at 15 cM and 45 cM positions with an equal effect.

raw data and takes advantage of the property of normal distribution, whereas the regression uses the squared phenotypic differences as the raw data. As a consequence, property of normal distribution is not utilized and also some information may have been lost.

The question now is how to calculate the proportion of genes shared IBD. If the parental mating type of the sibs are known, it is relatively easy to obtain the πs for all possible sib-pairs. However, this is difficult if the parental genotypes are unknown. However, given the allelic frequencies of the marker locus and assuming H-W equilibrium, the πs can be estimated. In both situations, the proportions of genes shared IBD between sibs are given in HASEMAN and ELSTON (1972). The proportions of genes IBD shared by half-sibs and other types of relatives in a complicated pedigree are usually estimated using a maximum likelihood approach (see AMOS *et al.* 1990). In this paper, we do not attempt to estimate the πs, rather we assume that the πs are known and thus focus on mapping procedures using these π values.

We have borrowed ZENG's (1994) idea of composite interval mapping for the multiple QTL model where we can treat genotypes of other markers outside the tested interval as fixed effects to control the genetic background. The problem here is that at linkage equilibrium the conditional probability of a QTL genotype given a marker genotype will remain unchanged in the whole population, leading to zero regression coefficients for all the marker genotypes. In fact, we must treat marker genotypes as nested within each family, which will increase the levels of the fixed effect. For instance, if all markers are fully informative, potentially there will be four genotypes per marker within each family. In the whole population, the number of levels for each marker may be up to $4n$, where n is the number of families. Unless there are a few families and each has large number of members, we can not simultaneously put all marker genotypes into the model. Therefore, the four marker approach in our multiple QTL model is appropriate.

The regression interval mapping of FULKER and CARDON (1994) still offers an advantage over the ML method in terms of computing speed. Hypothesis testing with a computationally fast algorithm can be easily accomplished by a recently developed permutation test method (CHURCHILL and DOERGE 1994) when an exact test is not available. In addition, the idea of composite interval mapping of ZENG (1994) and JANSEN (1994) can be directly adopted to FULKER and CARDON's (1994) mapping procedure by simply incorporating the IBD values of other important markers to control the genetic background and reduce the residual variance. To select important markers, one should consult JANSEN (1994).

One major advantage of using ML over regression is that the ML presents no difficulty for large sibships and irregular pedigree relationships. Although extension has been made to include large and complicated pedigrees by using weighted or generalized least squares methods (BLACKWELDER and ELSTON 1982; OLSON and WIJSMAN 1993), the methods still represent *ad hoc* approaches in terms of statistical testing, because the data (squared differences) are not normally distributed. In addition, the ML method presents no problem in incorporating fixed effects into the model to control the residual variance (SCHORK 1993). For such a mixed model, variance components are easily estimated using the well developed restricted maximum likelihood (REML) programs (*e.g.*, MEYER 1988). Upon incorporation of the interval mapping procedure into REML programs, QTL mapping in livestock and human populations will become routine. It is not feasible to use the ML method for extensive simulation studies due to prohibited computing load, but it can be widely used in real data analyses.

We are deeply indebted to Professors WILLIAM M. MUIR and ZHAO-BANG ZENG for helpful comments on an earlier version of the manuscript. We also want to acknowledge Professor SUBIR GHOSH for evaluating the ML algorithm presented in the manuscript. The current work was supported by National Institutes of Health grant GM-45344 and National Science Foundation grant BSR-910718.

LITERATURE CITED

AMOS, C. I., and R. C. ELSTON, 1989 Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genet. Epidemiol. **6:** 349–360.

AMOS, C. I., R. C. ELSTON, A. F. WILSON and J. E. VAILEY-WILSON, 1989 A more powerful robust sib-pair test of linkage for quantitative traits. Genet. Epidemiol. **6:** 435–449.

AMOS, C. I., D. V. DAWSON and R. C. ELSTON, 1990 The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. Am. J. Hum. Genet. **47:** 842–853.

ANDERSSON, L., C. S. HALEY, H. ELLEGREN, S. A. KNOTT, M. JOHANSSON *et al.*, 1994 Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science **263:** 1771–1774.

BLACKWELDER, W. C., and R. C. ELSTON, 1982 Power and robustness of sib-pair linkage tests and extension to larger sibships. Commun. Statist. Theor. Methods **11:** 449–484.

BISHOP, D. T., and J. A. WILLIAMSON, 1990 The power of identity-by-state methods for linkage analysis. Am. J. Hum. Genet. **46:** 254–265.

CHURCHILL, G. A., and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

DEGROOT, M. H., 1986 *Probability and Statistics*, Ed. 2. Addison-Wesley Publishing Company, Reading, MA.

FULKER, D. W., and L. R. CARDON, 1994 A sib-pair approach to interval mapping of quantitative trait loci. Am. J. Hum. Genet. **54:** 1092–1103.

GOLDGAR, D. E., 1990 Multipoint analysis of human quantitative genetic variation. Am. J. Hum. Genet. **47:** 957–967.

GOLDGAR, D. E., and R. S. ONIKI, 1992 Comparison of a multipoint identity-by-state method with parametric multipoint linkage analysis for mapping quantitative traits. Am. J. Hum. Genet. **50:** 598–606.

GRASER, H.-U., S. P. SMITH and B. TIER, 1987 A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. J. Anim. Sci. **64:** 1363–1370.

HALEY, C. S., and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

HALEY, C. S, S. A. KNOTT and J.-M. ELSEN, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics **136:** 1195–1207.

HASEMAN, J. K., and R. C. ELSTON, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2:** 3–19.

JANSEN, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

JANSEN, R. C., 1994 Controlling the Type I and Type II errors in mapping quantitative trait loci. Genetics **138:** 871–881.

KNOTT, S. A, and C. S. HALEY, 1992 Maximum likelihood mapping of quantitative trait loci using full-sib families. Genetics **132:** 1211–1222.

LANDER, E. S., and D. BOSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

MARTINEZ, O., and R. N. CURNOW, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85:** 480–488.

MEYER, K., 1988 DFREML—a set of programs to estimate variance components under an individual animal model. J. Dairy Sci. **71:** 33–34.

NELDER, J. A., and R. MEAD, 1965 A simplex method for function minimization. Comput. J. **7:** 308–313.

OLSON, J. M., and E. M. WIJSMAN, 1993 Linkage between quantitative trait and markers loci: methods using all relative pairs. Genet. Epidemiol. **10:** 87–102.

SAS Institute Inc., 1988 *SAS/STAT User's Guide*, Release 6.03. SAS Institute Inc., Cary, NC.

SCHORK, N. J., 1993 Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. Am. J. Hum. Genet. **53:** 1306–1393.

VAN ARENDONK, J. A. M., B. TIER and B. P. KINGHORN, 1994 Use of multiple genetic markers in prediction of breeding values. Genetics **137:** 319–329.

ZENG, Z.-B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA. **90:** 10972–10976.

ZENG, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: P. D. KEIGHTLEY

APPENDIX

A method of reparameterization is described here to ensure the simplex algorithm to search the unknowns in the appropriate solution space. Recall that the conditional covariance is $\mathrm{Cov}(y_{i1}, y_{i2} | \pi_{iL}\hat{\pi}_{iq}\pi_{iR}) = \sigma^2 r_i$, where

$$r_i = \pi_{iL} H_L^2 + \pi_{iq} h_g^2 + \pi_{iR} H_R^2 + \tfrac{1}{2} h_a^2$$

The permissible space of the heritabilities is

$$1 \geq H_L^2 \geq 0,\ 1 \geq h_g^2 \geq 0,\ 1 \geq H_R^2 \geq 0,\ 1 \geq h_a^2 \geq 0,$$

$$\text{and} \quad 1 \geq h^2 \geq 0$$

where $h^2 = H_L^2 + h_g^2 + H_R^2 + h_g^2$ is the overall heritability. These various heritabilities have to be searched within this space to guarantee a positive definite covariance matrix. Although we could borrow a technique of nonlinear programming from operations research that maximizes the likelihood function subject to these constraints, in this particular situation, we found that it is easier to use a method of reparameterization. The method of reparameterization is justified by the invariance property of the maximum likelihood estimators (DEGROOT 1986, p. 348). Defining $\gamma_1 = H_L^2/h^2$, $\gamma_2 = h_g^2/h^2$ and $\gamma_3 = H_R^2/h^2$ and $\gamma_4 = h_a^2/h^2$, we have

$$r_i = [\pi_{iL}\gamma_1 + \hat{\pi}_{iq}\gamma_2 + \pi_{iR}\gamma_3 + \tfrac{1}{2}\gamma_4] h^2$$

We now have five unknowns with a permissible space defined by

$$1 \geq h^2 \geq 0,\ \sum_{i=1}^{4} \gamma_i = 1 \quad \text{and} \quad \gamma_i \geq 0 \quad \text{for } i = 1, \ldots, 4$$

These newly defined terms can be reparameterized as

$$h^2 = \frac{\exp(z)}{1 + \exp(z)} \quad \text{and} \quad \gamma_i = \frac{\exp(x_i)}{\Sigma_{i=1}^{4} \exp(x_i)}$$

$$\text{for } i = 1, \ldots, 4$$

where z and x_i are any real numerals without any constraints. Instead of maximizing the likelihood function (L) with respect to γs and h^2, we will maximize L with respect to these xs and z. Let $\hat{x}_i$ and $\hat{z}$ denote the ML estimates of x_i and z, then the ML estimates of the original heritabilities are

$$\hat{h}^2 = \frac{\exp(\hat{z})}{1 + \exp(\hat{z})}, \quad \hat{H}_L^2 = \hat{h}^2 \frac{\exp(\hat{x}_1)}{\Sigma_{i=1}^{4} \exp(\hat{x}_i)},$$

$$\hat{h}_g^2 = \hat{h}^2 \frac{\exp(\hat{x}_2)}{\Sigma_{i=1}^{4} \exp(\hat{x}_i)}, \quad \hat{H}_R^2 = \hat{h}^2 \frac{\exp(\hat{x}_3)}{\Sigma_{i=1}^{4} \exp(\hat{x}_i)}$$

$$\text{and } \hat{h}_a^2 = \hat{h}^2 \frac{\exp(\hat{x}_4)}{\Sigma_{i=1}^{4} \exp(\hat{x}_i)}$$

In fact, the method of reparameterization presented here can be applied to any ML based methods for estimation of variance components. The usual practice that negative estimate of a variance component is replaced by zero (SAS Institute 1988, p. 967) is not correct, thus should not be used.