Genet Sel Evol (1996) 28, 479-490

Original article

Mapping quantitative trait loci in outcross populations via residual maximum likelihood. I. Methodology

FE Grignola [1], I Hoeschele [1*], B Tier [2]

[1] *Department of Dairy Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0315, USA;*
[2] *Animal Genetics and Breeding Unit, University of New South Wales, Armidale 2351, NSW, Australia*

(Received 18 March 1996; accepted 20 September 1996)

Summary – A residual maximum likelihood method, implemented with a derivative-free algorithm, was derived for estimating position and variance contribution of a single QTL together with additive polygenic and residual variance components. The method is based on a mixed linear model including random polygenic effects and random QTL effects, assumed to be normally distributed a priori. The method was developed for QTL mapping designs in livestock, where phenotypic and marker data are available on a final generation of offspring, and marker data are also available on the parents of the final offspring and on additional ancestors. The coefficient matrix of mixed model equations, required in the derivative-free algorithm, was derived from a reduced animal model linking single records of final offspring to parental polygenic and QTL effects. The variance–covariance matrix of QTL effects and its inverse were computed conditional on incomplete information from multiple linked markers. The inverse is computed efficiently for designs where each final offspring has a different dam and sires of the final generation have many genotyped progeny such that their marker linkage phase can be determined with a high degree of certainty. Linkage phases of ancestors of sires do not need to be known. Testing for a QTL at any position in the marker linkage group is based on the ratio of the likelihood estimating QTL variance to that with QTL variance set to zero.

quantitative trait loci / residual maximum likelihood / mapping

Résumé – La cartographie de locus de caractère quantitatif dans des populations en ségrégation à l'aide du maximum de vraisemblance résiduelle. I. Méthodologie. *Une méthode de maximum de vraisemblance résiduelle, appuyée sur un algorithme sans dérivation, a été établie pour estimer la position et la part de variance d'un locus de caractère quantitatif (QTL) et simultanément les variances polygéniques additives et résiduelles. La méthode est basée sur un modèle mixte linéaire incluant des effets*

* Correspondence and reprints.

polygéniques et de QTL, supposés a priori suivre une distribution normale. La méthode a été établie en vue de plans expérimentaux de cartographie de QTL chez les animaux, où des données phénotypiques et de marquage sont disponibles sur la génération des descendants, et des marquages connus chez les parents et des ancêtres supplémentaires. La matrice des coefficients des équations du modèle mixte, requise dans l'algorithme sans dérivation, a été déduite d'un modèle animal réduit qui relie les performances individuelles des descendants aux effets polygéniques ou de QTL parentaux. La matrice de variance-covariance des effets de QTL et son inverse ont été calculées conditionnellement à l'information incomplète relative à des ensembles de marqueurs liés. L'inverse est calculée efficacement pour des dispositifs où chaque descendant a une mère différente et les pères ont de nombreux descendants génotypés permettant de déterminer la phase des marqueurs liés avec un haut degré de certitude. Il n'est pas nécessaire de connaître la phase des marqueurs liés chez les ancêtres des pères. Le test de la position d'un marqueur au sein d'un groupe de liaison est basé sur le rapport de la vraisemblance correspondant à la variance QTL estimée, relative à une variance QTL nulle

locus de caractère quantitatif / maximum de vraisemblance résiduelle / cartographie

INTRODUCTION

Traditional methods for the statistical mapping of quantitative trait loci (QTL) include ANOVA and (multiple) linear regression (eg, Cowan et al, 1990; Weller et al, 1990; Haley et al, 1994; Zeng, 1994), maximum likelihood (ML) interval mapping (eg, Lander and Botstein, 1989; Knott and Haley, 1992), or a combination of ML and multiple regression interval mapping (eg, Zeng, 1994). These methods were developed mainly for line crossing and, hence, cannot fully account for the more complex data structures of outcross populations, eg, data on several families with relationships across families, unknown linkage phases in parents, unknown number of QTL alleles in the population, and varying amounts of data information on different QTLs or in different families. The gene effects near markers selected based on a linkage test tend to be overestimated increasingly with decreasing family size and true effect (Georges et al, 1995). Random treatment of QTL effects would cause shrinkage of estimates toward a prior mean in small families and for QTLs accounting only for a small portion of genetic variance.

Fernando and Grossman (1989) derived best linear unbiased prediction (BLUP) of QTL allelic effects, which are assumed to be normally distributed. For simple designs (eg, (grand)daughter designs with unrelated sires), BLUP reduces to random linear regression (Goddard, 1992). Fernando and Grossman (1989) showed how to obtain BLUP estimates of additive allelic effects (v) at a QTL linked to a single marker and of residual polygenic effects (u), assuming that all individuals in a population are genotyped and that markers are fully informative. Subsequent developments allowed for multiple linked markers with a QTL in each marker bracket (Goddard, 1992), for multiple unlinked markers each associated with a QTL (van Arendonk et al, 1994a), for incomplete marker information (Hoeschele, 1993; van Arendonk et al, 1994a; Wang et al, 1995), and for reductions in the number of equations by using a reduced animal model (RAM) (Cantet and Smith, 1991; Goddard, 1992), by including QTL gene effects only for genotyped animals and their tie ancestors (Hoeschele, 1993), or by estimating the sum of the effects at several unlinked, marked QTL (van Arendonk et al, 1994a). There are two linearly

equivalent (Henderson, 1985) animal models incorporating marker information, the first linking an individual's phenotype to both of its marked QTL allelic effects and to its polygenic effect (Fernando and Grossman, 1989), and the other linking phenotypes to the total additive effects and linking total additive effects to QTL effects via the genetic covariance matrix (Hoeschele, 1993).

All methods described above are concerned with the prediction of genetic effects and assume that the dispersion parameters are known. These parameters include the additive polygenic variance, the variance contributed by a QTL, the QTL position, and the residual variance. A first attempt to estimate these parameters by residual maximum likelihood (REML) was undertaken by van Arendonk et al (1994b) using a granddaughter design with unrelated sires and a single marker. These authors found that for this situation, QTL position and contribution to additive genetic variance were not separately estimable. Grignola et al (1994) showed that for the same type of design these parameters were estimable when performing interval mapping with flanking markers, known linkage phases in the sires and no relationships among sires.

Xu and Atchley (1995) performed interval mapping using maximum likelihood based on a mixed model with random QTL effects, but these authors fitted one additive genetic effect at the QTL rather than two allelic effects for each individual with variance–covariance matrix equal to a matrix of proportions of alleles identical-by-descent (IBD) shared by any two individuals at the QTL and assumed that this matrix was known. These authors applied their analysis to unrelated full-sib pairs.

In this paper, we (i) apply the theory of Wang et al (1995) for a single marker to compute the variance–covariance matrix among QTL effects conditional on incomplete information from multiple linked markers, (ii) use this covariance matrix in the estimation of position and variance contribution of a single QTL along with polygenic and residual variances and in testing for QTL presence in a marker linkage group via REML with a derivative-free algorithm, and (iii) include all known relationships between the parents (sires) of the final offspring in the analysis.

METHODOLOGY

Mixed linear model

The animal model including polygenic and QTL effects of Fernando and Grossman (1989) is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{T}\mathbf{v} + \mathbf{e}$$
$$\text{with } \mathrm{Var}(\mathbf{u}) = \mathbf{A}\sigma^2_{\mathrm{u}},\ \mathrm{Var}(\mathbf{v}) = \mathbf{Q}\sigma^2_{\mathrm{v}},\ \mathrm{Var}(\mathbf{e}) = \mathbf{R}\sigma^2_{\mathrm{e}} \qquad [1]$$

where $\mathbf{y}$ is an $N \times 1$ vector of phenotypes, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{X}$ is a design/covariate matrix relating $\boldsymbol{\beta}$ to $\mathbf{y}$, $\mathbf{u}$ is an $n \times 1$ vector of residual additive (polygenic) effects, $\mathbf{Z}$ is an incidence matrix relating records in $\mathbf{y}$ to animals, $\mathbf{v}$ is a $2n \times 1$ vector of QTL allelic effects, $\mathbf{T}$ is an incidence matrix relating each animal to its two QTL alleles, $\mathbf{e}$ is a vector of residuals, $\mathbf{A}$ is the additive genetic relationship matrix, σ^2_{u} is the polygenic variance, $Q\sigma^2_{\mathrm{v}}$ is the variance–covariance matrix of the QTL allelic effects conditional on marker information, σ^2_{v} is half the

additive genetic variance explained by the QTL (also referred to as the QTL allelic variance), $\mathbf{R}$ is a known diagonal matrix, and σ_e^2 is the residual variance.

Matrix $\mathbf{Q}$ depends on one unknown parameter, the map position of the QTL relative to the origin of the marker linkage group (d_Q). For notational convenience, this dependency is suppressed in model [1] and below. Parameters related to the marker map (marker distances and allele frequencies) are assumed to be known.

The model is parameterized in terms of the unknown parameters heritability ($h^2 = \sigma_a^2/\sigma_p^2$) with σ_a^2 being additive genetic and σ_p^2 phenotypic variance, fraction of the additive genetic variance explained by the QTL allelic variance or half of the additive variance due to the QTL ($v^2 = \sigma_v^2/\sigma_a^2$), residual variance σ_e^2, and QTL location d_Q.

Let there be phenotypes only on nonparents or final offspring which have single records. Furthermore, recurrence equations linking u and v effects of nonparents (n) to those of parents (p) are

$$\mathbf{u}_n = \mathbf{W}\mathbf{u}_p + \mathbf{m},\ \mathbf{v}_n = \mathbf{F}\mathbf{v}_p + \boldsymbol{\epsilon} \qquad [2]$$

where the matrix $\mathbf{W}$ consists of rows with zero, one or two elements equal to 0.5 for none, one or two parents known, respectively, and each row of the matrix $\mathbf{F}$ contains up to four nonzero coefficients explained below. With single records, $\mathbf{Z} = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. Then, model [1] can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u}_p + \mathbf{T}\mathbf{F}\mathbf{v}_p + \mathbf{m} + \mathbf{P}\boldsymbol{\epsilon} + \mathbf{e}$$
$$\mathrm{Var}(\mathbf{u}_p) = \mathbf{A}_p\sigma_u^2,\ \mathrm{Var}(\mathbf{v}_p) = \mathbf{Q}_p\sigma_v^2,\ \mathrm{Var}(\mathbf{m}) = \boldsymbol{\Delta}_u\sigma_u^2,\ \mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Delta}_v\sigma_v^2,\ \mathrm{Var}(\mathbf{e}) = \mathbf{R}\sigma_e^2 \qquad [3]$$

The reduced animal model is obtained from [3] by combining the last three terms into the residual. Mixed model equations (MME) for the RAM (Cantet and Smith, 1991) can be formed based on the RAM directly or by first forming MME based on [3] and subsequently absorbing the equations in $\boldsymbol{\epsilon}$ and $\mathbf{m}$. The resulting MME for the RAM are

$$\begin{bmatrix} \mathbf{X}'\mathbf{D}_{uv}\mathbf{X} & \mathbf{X}'\mathbf{D}_{uv}\mathbf{W} & \mathbf{X}'\mathbf{D}_{uv}\mathbf{T}\mathbf{F} \\ \mathbf{W}'\mathbf{D}_{uv}\mathbf{X} & \mathbf{W}'\mathbf{D}_{uv}\mathbf{W} + \mathbf{A}^{-1}k_u & \mathbf{W}'\mathbf{D}_{uv}\mathbf{T}\mathbf{F} \\ \mathbf{F}'\mathbf{T}'\mathbf{D}_{uv}\mathbf{X} & \mathbf{F}'\mathbf{T}'\mathbf{D}_{uv}\mathbf{W} & \mathbf{F}'\mathbf{T}'\mathbf{D}_{uv}\mathbf{T}\mathbf{F} + \mathbf{Q}^{-1}k_v \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}_p \\ \mathbf{v}_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{D}_{uv}\mathbf{y} \\ \mathbf{W}'\mathbf{D}_{uv}\mathbf{y} \\ \mathbf{F}'\mathbf{T}'\mathbf{D}_{uv}\mathbf{y} \end{bmatrix} \qquad [4]$$

$$\text{where } \mathbf{D}_{uv} = \mathbf{D}_v - \mathbf{D}_v\left(\mathbf{D}_v + \boldsymbol{\Delta}_u^{-1}k_u\right)^{-1}\mathbf{D}_v = \left(\mathbf{D}_v^{-1} + \boldsymbol{\Delta}_u\frac{1}{k_u}\right)^{-1}$$

$$\mathbf{D}_v = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{T}\left(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Delta}_v^{-1}k_v\right)^{-1}\mathbf{T}'\mathbf{R}^{-1} = \left(\mathbf{R} + \mathbf{T}\boldsymbol{\Delta}_v\mathbf{T}'\frac{1}{k_v}\right)^{-1}$$

$$k_u = \frac{\sigma_e^2}{\sigma_u^2} \quad \text{and} \quad k_v = \frac{\sigma_e^2}{\sigma_v^2}$$

It can be easily verified that matrix $\mathbf{D}_v$ is diagonal even if $\boldsymbol{\Delta}_v$ is not, ie, $\mathbf{T}\boldsymbol{\Delta}_v\mathbf{T}'$ is always diagonal. Inbreeding and unknown parental origin of marker alleles can

give rise to some nonzero offdiagonal elements in $\mathbf{\Delta}_v$ (Hoeschele, 1993; Wang et al, 1995). With $\mathbf{D}_v$ diagonal, matrix $\mathbf{D}_{uv}$ is also diagonal, hence, the MME are easily computed.

REML analysis

The REML analysis was performed by maximizing the likelihood of error contrasts (LEC) (Patterson and Thompson, 1971) with respect to the parameters h^2, v^2, σ_e^2, and d_Q. The LEC was obtained under the assumption of a joint multivariate normal distribution of $\mathbf{y}$, $\mathbf{u}$, and $\mathbf{v}$. For the full animal model (AM), the logarithm of the LEC (LLEC) can be expressed as (Meyer, 1989):

$$\text{LLEC}_{AM} = \log L(\mathbf{y};\boldsymbol{\theta}) = \text{const} - 0.5 \log|\mathbf{G}| - 0.5(N - NF - NR) \log(\hat{\sigma}_e^2) - 0.5 \log|\mathbf{C}| - 0.5\mathbf{y}'\mathbf{P}\mathbf{y}\hat{\sigma}_e^{-2} \quad [5]$$

$$\text{where } \mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$$

$$\hat{\sigma}_e^2 = \frac{\mathbf{y}'\mathbf{P}\mathbf{y}}{N - NF}, \quad \mathbf{V} = \text{Var}(\mathbf{y})\frac{1}{\sigma_e^2}$$

where $\boldsymbol{\theta}$ is the vector of parameters, $\mathbf{G}$ is the variance–covariance matrix of the random effects (here, $\mathbf{u}$ and $\mathbf{v}$), NF = rank($\mathbf{X}$), NR = dimension($\mathbf{G}$), and $\mathbf{C}$ is the coefficient matrix of the MME for the AM (model [1] or [3]) reparameterized to full rank and with σ_e^2 factored out. The estimate of σ_e^2 maximizes the likelihood for a given set of values for the other parameters (Graser et al, 1987). The terms $\mathbf{y}'\mathbf{P}\mathbf{y}$ and $\log|\mathbf{C}|$ are computed as in Meyer (1989) via Gaussian elimination applied to the augmented MME, and $\log|\mathbf{G}|$ is obtained as $-\log|\mathbf{G}^{-1}|$ with $\mathbf{G}^{-1}$ computed directly. In the following, it is shown how to compute the AM likelihood in [5] when working with MME for the RAM.

When equation [5] is applied directly to the RAM, the result is

$$\text{LLEC}_{RAM} \propto -0.5 \log|\mathbf{G}_{RAM}| - 0.5(N - NF - NR_{RAM}) \log(\hat{\sigma}_e^2) - 0.5 \log|\mathbf{C}_{RAM}| - 0.5\mathbf{y}'\mathbf{P}\mathbf{y}\hat{\sigma}_e^{-2} \quad [6]$$

where all parts different from the AM LLEC are subscripted RAM. Let $\mathbf{G}$ be the variance–covariance matrix of the genetic effects $(\mathbf{u}, \mathbf{v})$ of the parents and of the Mendelian sampling effects for $\mathbf{u}$ and $\mathbf{v}$ of the nonparents or finals. Let $\mathbf{G}$ be partitioned accordingly. Then

$$\log|\mathbf{G}^{-1}| = \log\begin{vmatrix} \mathbf{G}_p^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}^{-1} \end{vmatrix} = \log|\mathbf{G}_p^{-1}| + \log|\mathbf{\Delta}^{-1}| \quad [7]$$

$$\text{where } \mathbf{G}_p = \begin{bmatrix} \mathbf{A}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_p \end{bmatrix}, \quad \mathbf{\Delta} = \begin{bmatrix} \mathbf{\Delta}_u & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}_v \end{bmatrix}$$

where $\mathbf{\Delta}$ is blockdiagonal with blocks of size $\leqslant 2$. Similarly, partition the coefficient matrix of the MME for model [3], $\mathbf{C}$, according to all other (1: $\boldsymbol{\beta}$, $\mathbf{u}_p$, $\mathbf{v}_p$) and Mendelian sampling effects (2: $\mathbf{m}$, $\boldsymbol{\epsilon}$). Then,

$$\log|\mathbf{C}| = \log|\mathbf{C}_{22}| + \log|\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}| \quad [8]$$

where $\mathbf{C}_{22}$ is diagonal or blockdiagonal with blocks of size $\leqslant 2$. Hence, the RAM LLEC can easily be modified to yield the AM LLEC, or

$$\text{LLEC}_{\text{RAM}} \propto \text{LLEC}_{\text{RAM}} - 0.5 \log|\mathbf{\Delta}| - 0.5 \log|\mathbf{C}_{22}| + 0.5\ (\text{NR} - \text{NR}_{\text{RAM}}) \log(\partial_{\text{e}}^2) \quad [9]$$

where NR is total number of random genetic effects while NR_{RAM} is number of genetic effects pertaining to parents.

The analysis was conducted in the form of interval mapping as in Xu and Atchley (1995), where d_{Q} was fixed at a number of successive positions (every centimorgan) along the chromosome, and at each position the likelihood was maximized with respect to h^2, v^2, and σ_{e}^2.

Calculation of $\mathbf{Q}_{\text{p}}^{-1}$ *and* $\mathbf{\Delta}_{\text{v}}$

These matrices were computed by applying the theory presented in Wang et al (1995) to marker information consisting of multiple linked rather than single markers. At a given QTL position, different markers were allowed to flank the QTL in different families due to some parents being homozygous at the closest flanking markers.

Notation

Let Q_{i}^k denote QTL allele k ($k = 1, 2$) in individual i and v_i^k the additive effect of this allele. Let $\equiv$ denote IBD, let $\Leftarrow$ stand for 'inherited from', let $\mathbf{G}_{\text{obs}}$ represent the marker information observed on the pedigree, and let M_i^m denote a possible marker haplotype ($\shortparallel$) of individual i at the closest pair of marker loci bracketing the QTL for which the parent of i is heterozygous. Furthermore, let $\mathbf{M}$ be a set of complete multi-locus marker genotypes for the entire pedigree. Finally, p denotes parent (p = s, d), s sire, d dam, and L_{p} denotes the linkage phase of the alleles at the narrowest marker bracket for which parent p is heterozygous.

Variance–covariance matrix of v effects $\mathbf{Q}_{\text{p}}$

In the presence of missing marker data and/or unknown linkage phases for parents, the variance–covariance matrix of the v effects is of the form

$$\mathbf{Q}_{\mathbf{G}_{\text{obs}}} = \sum_{\mathbf{M}} \Pr(\mathbf{M}|\mathbf{G}_{\text{obs}})\mathbf{Q}_{\text{M}} \quad [10]$$

where $\mathbf{Q}_{\text{M}}$ is conditional on a particular set of multi-locus marker genotypes ($\mathbf{M}$). Equation [10] was given in Hoeschele (1993) and in Wang et al (1995). The calculation of [10] is computationally very demanding for large pedigrees. The probability of a QTL allele in individual i being IBD to a QTL allele in individual j (with j not being a direct descendant of i) in general cannot be computed recursively using IBD probabilities pertaining to the alleles in i and the parents of j when parental marker genotypes and/or linkage phases are unknown (Wang et al, 1995), hence there is no simple method to compute the inverse directly. A method for

computing the inverse, which is more efficient than standard inversion, was derived by Van Arendonk et al (1994a).

The variance-covariance matrix in [10] can, however, be computed by using Monte Carlo. The Monte-Carlo approximation of [10] is

$$\mathbf{Q}_{\mathbf{G}_{\text{obs}}} \approx \frac{1}{S} \sum_{k=1}^{k=s} \mathbf{Q}_{\mathbf{M}_k} \qquad [11]$$

where $\mathbf{M_k}$ is a particular realization of $\mathbf{M}$ from the probability distribution of $\mathbf{M}$ given $\mathbf{G}_{\text{obs}}$, and S is sample size. Note that [11] yields the exact variance–covariance matrix if sample size S is large. Samples from this distribution can be obtained by Gibbs sampling, which was implemented using blocking of the genotypes of parents and final offspring (finals) as in Janss et al (1995). For a half-sib design (daughter or granddaughter design) with large family sizes (eg, 50–100) and no relationships among final offspring (daughters or sons) through dams, the linkage phases of the parents of final offspring are 'known', as always or most frequently (near 100%) the correct phase is sampled. Then, the inverse of the variance–covariance matrix of the QTL effects can be computed exactly (up to Monte-Carlo error due to use of [11]) as follows. Equation [11] is employed to compute the submatrix pertaining to QTL effects of parents of finals and ancestors using marker information on the entire pedigree including final offspring. This submatrix is then inverted, and contributions of final offspring, computed with known parental linkage phases, are added into the inverse. Note that in the RAM in [4], offspring contributions appear in the least-squares part of the MME rather than in the inverse variance–covariance matrix of the QTL effects.

Recurrence equations for v effects

Recurrence equations for the v effects of the finals were required to compute the elements of $\mathbf{F}$ and $\mathbf{\Delta}_{\text{v}}$ in [4]. The general recurrence equation for a QTL effect is

$$v_{\text{i}}^k = \sum_{\text{p=s,d}} \sum_{l=1,2} \Pr(Q_i^k \equiv Q_{\text{p}}^l | \mathbf{G}_{\text{obs}}) v_{\text{p}}^l + \epsilon_i^k \qquad [12]$$

where

$$\Pr(Q_i^k \equiv Q_{\text{p}}^l | \mathbf{G}_{\text{obs}}) = \sum_{\text{L}_\text{p}} \sum_{m} \Pr(\text{L}_\text{p}, M_i^m \Leftarrow p | \mathbf{G}_{\text{obs}}) \Pr(Q_i^k \equiv Q_{\text{p}}^l | M_i^m \Leftarrow \text{p}, \text{L}_\text{p}) \qquad [13]$$

The most likely linkage phase is assumed to be the true phase for the parents of final offspring. This assumption reduces the joint probability of parental linkage phase and offspring haplotype in [13] to the probability of the marker haplotype of an offspring. This probability is computed using the parental phase and the marker genotypes of an offspring at all linked markers. Alternatively, [13] could be used when parental linkage phases are not known by computing the joint probability of parent linkage phase and offspring haplotype for each interval as a frequency count across all Gibbs cycles after burn-in, using information from the entire marker

linkage group and from all relatives in the pedigree. However, this approach would only be an approximation to calculating the variance–covariance matrix and its inverse based on [10] for the entire pedigree including the final offspring.

In [13], the $\Pr(Q_i^k \equiv Q_p^l | M_i^m \Leftarrow \mathrm{p}, \mathrm{L_p})$ are $t_{11} = (1-r_\mathrm{L})(1-r_\mathrm{R})/(1-r_\mathrm{M})$ and $t_{12} = r_\mathrm{L} r_\mathrm{R}/(1-r_\mathrm{M})$ if M_i^m is a nonrecombinant haplotype, or $t_{21} = (1-r_\mathrm{L})r_\mathrm{R}/r_\mathrm{M}$ and $t_{22} = r_\mathrm{L}(1-r_\mathrm{R})/r_\mathrm{M}$ if M_i^m is recombinant, where r_M is recombination rate for the marker bracket, $r_\mathrm{L}(r_\mathrm{R})$ is recombination rate between the QTL and the left (right) marker, and Haldane's no interference map function is employed. Here, we allow for double recombination while Goddard (1992) assumed it to be zero.

QTL alleles in final offspring are identified by parental origin, ie, the two QTL alleles in an offspring are distinguished as the allele inherited from the sire (s) and the allele coming from the dam (d). This definition can be employed even if the parental origins of the alleles at the flanking markers are unknown, but it can be used only in the final generation. For illustration, consider a single parent p (here, p = s = sire) with genotype 12/12, linkage phase 1 − 1, and the worst case of an offspring with genotype 12/12 (inheritance unknown at both flanking markers). The possible marker haplotypes inherited from p are 1 − 1, 1 − 2, 2 − 1, and 2 − 2. Then, if the QTL alleles in i are identified by the alleles at the left marker (1,2)

$$v_i^1 = [\Pr(1-1 \Leftarrow \mathrm{s})\, t_{11} + \Pr(1-2 \Leftarrow \mathrm{s})\, t_{21}]v_s^1 + [\Pr(1-1 \Leftarrow \mathrm{s})\, t_{12} + \Pr(1-2 \Leftarrow \mathrm{s})\, t_{22}]v_s^2 + [\Pr(2-1 \Leftarrow \mathrm{s}) + \Pr(2-2 \Leftarrow \mathrm{s})]0.5(v_d^1 + v_d^2) + \epsilon_i^1$$

$$v_i^2 = [\Pr(2-1 \Leftarrow \mathrm{s}) t_{22} + \Pr(2-2 \Leftarrow \mathrm{s})\, t_{12}]v_s^1 + [\Pr(2-1 \Leftarrow \mathrm{s})\, t_{21} + \Pr(2-2 \Leftarrow \mathrm{s})\, t_{11}]v_s^2 + [\Pr(1-1 \Leftarrow \mathrm{s}) + \Pr(1-2 \Leftarrow \mathrm{s})]0.5(v_d^1 + v_d^2) + \epsilon_i^2$$

whereas if the QTL alleles in i are identified by parental origin,

$$v_i^1 = [\Pr(1-1 \Leftarrow \mathrm{s})\, t_{11} + \Pr(1-2 \Leftarrow \mathrm{s})\, t_{21} + \Pr(2-1 \Leftarrow \mathrm{s})\, t_{22} + \Pr(2-2 \Leftarrow \mathrm{s})\, t_{12}]v_s^1 + [\Pr(1-1 \Leftarrow \mathrm{s})\, t_{12} + \Pr(1-2 \Leftarrow \mathrm{s})\, t_{22} + \Pr(2-1 \Leftarrow \mathrm{s})\, t_{21} + \Pr(2-2 \Leftarrow \mathrm{s})\, t_{11}]v_s^2 + \varepsilon_i^1$$

$$v_i^2 = 0.5v_d^1 + 0.5v_d^2 + \epsilon_i^2$$

Note that summing the v_i^1 and v_i^2 equations yields the same result for both QTL identifications. Note also that the advantage of the identification by parental origin is that only v_i^2 is linked to the v effects of the dam (d), instead of linking both v_i^1 and v_i^2 to the dam effects requiring to include dam effects in the MME.

Hypothesis test

The likelihood under the null hypothesis is evaluated at $v^2 = 0$. The distribution of the likelihood ratio statistic is not known exactly, regardless of the method used to locate QTL (Churchill and Doerge, 1994). For the null hypothesis postulating the absence of a QTL in a particular interval rather than in the entire genome, Xu and Atchley (1995) found the distribution to be in between two chi-square distributions with degrees of freedom of one and two, respectively. Several factors may influence the distribution of a test statistic for QTL presence, eg, the length of the genome, the marker density, the extent to which marker data are missing,

segregation distortion, and the distribution of the phenotypes. Self and Liang (1987) derived analytical results for the asymptotic distribution of the likelihood ratio statistic for cases where the true parameter value may be on the boundary of the parameter space. However, with finite sample sizes and several factors influencing the distribution of the statistic, it is questionable whether their results can be utilized in QTL mapping.

When analyzing real data, the threshold value for significance can be determined empirically using data permutation (Churchill and Doerge, 1994). To obtain the threshold value for a genome-wide search, in the order of 10 000 to 100 000 permutations are necessary. As these computations are unfeasible with the method presented here (see the companion paper by Grignola et al, 1996), one may resort to estimating thresholds for a number of less stringent significance levels and obtain the desired threshold by extrapolation (Uimari et al, 1996b).

CONCLUSIONS

The REML analysis described in this paper may be a useful alternative to other methods for the statistical mapping of QTL. The REML method is generally known to be quite robust to deviations from normality. When applied to QTL mapping, the REML analysis requires fewer parametric assumptions than ML (eg, Weller, 1986) and Bayesian analyses (Hoeschele et al, 1996; Thaller and Hoeschele, 1996a,b; Uimari et al, 1996a) postulating a biallelic QTL with unknown gene frequency and substitution effect.

While Xu and Atchley (1995) estimate QTL, polygenic and residual variances by ML, we perform REML estimation. While REML should be preferred over ML in the presence of many fixed effects relative to the number of observations (Patterson and Thompson, 1971), a model for the analysis of QTL mapping experiments may only need to include an overall mean. In this case, the difference between the ML and REML analyses is negligible.

As the true nature of QTLs is unknown, it is important to evaluate the performance of this REML analysis and of other methods with data simulated under different genetic models (eg, biallelic and multiallelic QTL models). In a companion paper (Grignola et al, 1996), we apply the REML analysis to granddaughter designs simulated with different models for the additive variance at the QTL. Hoeschele et al (1996) apply Bayesian analyses based on biallelic and multiallelic QTL models to data simulated under both models.

The REML analysis incorporates an expected variance–covariance matrix of the QTL allelic effects, which is equal to a weighted average of variance–covariance matrices conditional on all possible sets of multi-locus marker genotypes given the observed marker data. Schork (1993) alternatively formulated a likelihood for a mixture distribution which is a weighted average of REML likelihoods conditional on all possible sets of multi-locus marker genotypes given the observed marker data. He pointed out, however, that simulation results indicated that his modification may lead to a loss of power. In both approaches, the one considered in this paper and in equivalent form by Xu and Atchley (1995), and the approach of Schork (1993), probabilities of multi-locus marker genotypes are computed from the observed marker information. However, if markers are linked to QTLs,

phenotypes also contain information about marker genotypes, and this information is ignored here (Van Arendonk, personal communication). In this regard, the REML analysis can be viewed as an approximation to the Bayesian analysis based on a multiallelic QTL model with QTL variance and allelic effects having a prior normal distribution (Hoeschele et al, 1996). The Bayesian analysis takes into account the joint distribution of the QTL and marker genotypes conditional on the phenotypic information.

We are currently extending our REML analysis to account for multiple linked QTLs. One way of approaching this problem was presented by Xu and Atchley (1995) and consisted of fitting variances associated with next-to-flanking markers. Disadvantages of this approach are that it is approximate as effects associated with marker alleles identified within founders erode over generations, and that it requires many additional parameters when the marker polymorphism is limited, causing the flanking and next-to-flanking markers to differ among families.

Finally, we plan to extend the REML analysis to other designs (eg, full-sib designs), where the current computation of the inverse of the variance–covariance matrix becomes approximate due to uncertain linkage phases in parents of final offspring, and other ways of computing this inverse exactly (eg, Van Arendonk et al, 1994a) will be implemented.

ACKNOWLEDGMENTS

This research was supported by Award No 92-01732 of the National Research Initiative Competitive Grants Program of the US Department of Agriculture and by the Holstein Association USA. I Hoeschele acknowledges financial support from the European Capital and Mobility fund while on research leave at Wageningen University, the Netherlands. B Tier achnowledges financial support from the Australian Department of Industry, Training, and Regional Development while on research leave at Virginia Polytechnic Institute and State University.

REFERENCES

Cantet RJC, Smith C (1991) Reduced animal model for marker assisted selection using best linear unbiased prediction. *Genet Sel Evol* 23, 221-233

Churchill G, Doerge R (1994) Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963-971

Cowan CM, Dentine MR, Ax RL, Schuler LA (1990) Structural variation around prolactin gene linked to quantitative traits in an elite Holstein family. *Theor Appl Genet* 79, 577-582

Fernando RL, Grossman M (1989) Marker-assisted selection using best linear unbiased prediction. *Genet Sel Evol* 21, 467-477

Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, Pasquino AT, Sargeant LS, Sorensen A, Steele MR, Zhao X, Womack JE, Hoeschele I (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* 139, 907-920

Goddard M (1992) A mixed model for analyses of data on multiple genetic markers. *Theor Appl Genet* 83, 878-886

Graser HU, Smith SP, Tier B (1987) A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *J Anim Sci* 64, 1363-1370

Grignola FE, Hoeschele I, Meyer K (1994) Empirical best linear unbiased prediction to map QTL. *Proc 5th World Congr Genet Appl Livest Prod* 21, 245-248

Grignola FE, Hoeschele I, Thaller G (1996) Mapping quantitative trait loci in outcross populations via residual maximum likelihood. II. A simulation study. *Genet Sel Evol* 28, 491-504

Haley CS, Knott SA, Elsen JM (1994) Mapping quantitative trait loci in crosses between outbred lines using least-squares. *Genetics* 136, 1195-1207

Henderson CR (1985) Equivalent linear models to reduce computations. *J Dairy Sci* 68, 2267-2277

Hoeschele I (1993) Elimination of quantitative trait loci equations in an animal model incorporating genetic marker data. *J Dairy Sci* 76, 1693-1713

Hoeschele I, Uimari P, Grignola FE, Zhang Q, Gage KM (1996) Statistical mapping of polygene loci in livestock. In: *Proc Int Biometric Soc* (in press)

Janss LLG, Thompson R, van Arendonk JAM (1995) Application of Gibbs sampling for inference in a mixed major gene–polygenic inheritance model in animal populations. *Theor Appl Genet* 91, 1137-1147

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

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

Meyer K (1989) Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genet Sel Evol* 21, 317-340

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545-554

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

Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio statistics under nonstandard conditions. *J Am Stat Assoc* 82, 605-610

Thaller G, Hoeschele I (1996a) A Monte-Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci: I. Methodology. *Theor Appl Genet* 92 (in press)

Thaller G, Hoeschele I (1996b) A Monte-Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci: I. A simulation study. *Theor Appl Genet* 92 (in press)

Uimari P, Thaller G, Hoeschele I (1996a) The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* 143, 1831-1842

Uimari P, Zhang Q, Grignola F, Hoeschele I, Thaller G (1996b) Analysis of QTL workshop I granddaughter design data using least-squares, residual maximum likelihood and Bayesian methods. *J Quantitative Trait Loci* (in press)

van Arendonk JAM, Tier B, Kinghorn BP (1994a) Use of multiple genetic markers in prediction of breeding values. *Genetics* 137, 319-329

van Arendonk JAM, Tier B, Kinghorn BP (1994b) Simultaneous estimation of effects of unlinked markers and polygenes on a trait showing quantitative genetic variation. Proc 17th Int Congr Genetics, Birmingham, UK, p 192

Wang T, Fernando RL, van der Beek S, M Grossman (1995) Covariance between relatives for a marked quantitative trait locus. *Genet Sel Evol* 27, 251-274

Weller JI (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* 42, 627-640

Weller JI, Kashi Y, Soller M (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *J Dairy Sci* 73, 2525-2537

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

Zeng ZB (1994) Precision mapping of quantitative trait loci. *Genetics* 136, 1457-1468