Genet Sel Evol (1993) 25, 537-555

Original article

Genetic evaluation for a quantitative trait controlled by polygenes and a major locus with genotypes not or only partly known

A Hofer [1], BW Kennedy [2]

[1] *Department of Animal Sciences, Federal Institute of Technology (ETH), CH-8092 Zürich, Switzerland;*
[2] *Centre for Genetic Improvment of Livestock, University of Guelph, Guelph, Ontario, N1G 2W1, Canada*

(Received 4 March 1992; accepted 5 August 1993)

Summary – For a quantitative trait controlled by polygenes and a major locus with 2 alleles, equations for the maximum likelihood estimation of major locus genotype effects and polygenic breeding values, as well as major allele frequency and major locus genotype probabilities, were derived. Because the resulting expressions are computationally untractable for practical application, possible approximations were compared with 2 other procedures suggested in the literature using stochastic computer simulation. Although the frequency of the favourable allele was seriously underestimated when major locus genotypes were entirely unknown, the proposed method compares favourably with the 2 other procedures under certain conditions. None of the procedures compared can satisfactorily separate major genotypic effects from polygenic effects. However, the proposed method has some potential for improvement.

major locus / genetic evaluation / segregation analysis

Résumé – Évaluation génétique pour un caractère quantitatif contrôlé par des polygènes et un locus majeur à génotypes inconnus ou seulement partiellement connus. *Pour un caractère contrôlé par des polygènes et un locus majeur à 2 allèles, les équations pour l'estimation du maximum de vraisemblance des effects génotypiques au locus majeur et des valeurs génétiques polygéniques ont été dérivées, permettant aussi d'estimer la fréquence de l'allèle majeur et les probabilités des génotypes à ce locus. Les expressions obtenues étant incalculables en pratique, des approximations possibles ont été comparées par simulation stochastique à 2 autres procédures proposées dans la littérature. Bien que la fréquence de l'allèle favorable soit sérieusement sous-estimée lorsque les génotypes au locus majeur sont entièrement inconnus, la méthode proposée a quelques avantages sur les 2 autres procédés sous certaines conditions. Aucune des procédures comparées n'est*

satisfaisante pour séparer l'effet des génotypes majeurs des effets polygéniques. Cependant, la méthode proposée est susceptible d'être améliorée.

locus majeur / évaluation génétique / analyse de ségrégation

INTRODUCTION

Statistical methods based on the infinitesimal model, the assumption of many unlinked loci all with small effects controlling quantitative traits, have been successfully applied in animal breeding. An increasing number of studies, however, have reported single loci having large effects on quantitative traits. Such loci are referred to as major loci. Examples are the prolactin (Cowan *et al*, 1990) and the weaver loci (Hoeschele and Meinert, 1990) in dairy cattle, and the halothane sensitivity locus (Eikelenboom *et al*, 1980) and a locus acting on "Napole" yield (Le Roy *et al*, 1990), a pork quality trait, in pigs. Only in the case of the halothane locus has the responsible gene been identified and procedures for its genotyping become available (MacLennan and Phillips, 1992).

There is no difficulty with genetic evaluation for traits controlled by a major locus and polygenes when major locus genotypes are known. A fixed major locus effect has to be added to the linear model and major locus effects and polygenic breeding values can be estimated by the usual mixed model equations (Kennedy *et al*, 1992). When genotypes are unknown, however, satisfactory statistical methods are still lacking. Selection decisions could possibly be based on animal models that include the major locus effects in the polygenic part of the model. In cases where the allele has some positive effect on 1 trait but negative effects on others, it would be desirable to have separate estimates of the major locus and polygenic effects available. The 2 estimates would then be combined according to the breeding objective. Because genotyping of all the animals of a population is likely to be too expensive if at all possible, statistical methods are required that estimate major locus genotype effects as well as polygenic effects and major locus genotype probabilities for each candidate.

Such a method was first proposed in human genetics by Elston and Stewart (1971). The unknown parameters of the model are estimated by maximizing the likelihood of the data. For models with both major locus and polygenic effects exact calculations are very expensive and become unfeasible for pedigrees with more than $\sim$ 15 individuals. Several studies compared the power of different approximations of the likelihood function to detect a major locus in half-sib family structures in animal breeding data (Le Roy *et al*, 1989; Elsen and Le Roy, 1989; Knott *et al*, 1992a). Hoeschele (1988) developed an iterative procedure to estimate major locus genotype probabilities and effects as well as polygenic breeding values. The equations produced for the estimation of genotype probabilities were derived for simple population structures and were based on an approximation of the likelihood function. Kinghorn *et al* (1993) used the iterative algorithm of van Arendonk *et al* (1989) to estimate genotype probabilities and estimated genotype effects by

regression on genotype probabilities. A method was proposed to correct for the bias inherent in such analyses.

The objectives of this study were: i) to derive exact maximum likelihood equations to estimate major locus genotype probabilities and effects for a quantitative trait with mixed major locus and polygenic inheritance without any restrictions on population structure; ii) to examine possible approximations; and iii) to compare these approximations with the methods of Hoeschele (1988) and Kinghorn *et al* (1993) by stochastic computer simulation.

METHODS

Model

Consider a quantitative trait which is controlled by 1 autosomal major locus with 2 alleles, A and a, and many other unlinked loci with alleles of small effects. Mendelian segregation is assumed for all alleles at all loci. The allele with the major effect, A, has a frequency of p in the base population, which is assumed to be unselected, not inbred and in Hardy-Weinberg and gametic equilibria. In the base population the 3 possible genotypes at the major locus (AA, Aa and aa), which will be denoted as 1, 2 and 3 throughout this paper, are therefore expected to occur in frequencies of p^2, $2p(1-p)$ and $(1-p)^2$, respectively. Because genotyping of animals might be impossible or too expensive, we assume for the moment that the genotypes at the major locus are not known. With 1 observation per animal the following mixed linear model can be formulated:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZTg} + \mathbf{Za} + \mathbf{e}, \qquad [1]$$

where $\mathbf{y}$ = observation vector
$\mathbf{b}$ = vector of non-genetic fixed effects
$\mathbf{g}$ = vector of fixed major locus genotype effects $[g_1\ g_2\ g_3]'$
$\mathbf{a}$ = vector of random polygenic breeding values
$\mathbf{e}$ = vector of random errors
$\mathbf{X}$,$\mathbf{Z}$ = known incidence matrices
$\mathbf{T}$ = unknown incidence matrix indicating true major locus genotypes of all the animals in the population

The expectation and variance of the random variables are assumed to be:

$$\mathrm{E}\begin{bmatrix}\mathbf{y}\\ \mathbf{a}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{Xb}+\mathbf{ZTg}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}, \quad \mathrm{Var}\begin{bmatrix}\mathbf{y}\\ \mathbf{a}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{V} & \mathbf{ZA}\lambda^{-1} & \mathbf{I}\\ \mathbf{AZ}'\lambda^{-1} & \mathbf{A}\lambda^{-1} & \mathbf{0}\\ \mathbf{I} & \mathbf{0} & \mathbf{I}\end{bmatrix}\cdot\sigma_e^2$$

where $\mathbf{V} = \mathbf{ZAZ}'\lambda^{-1} + \mathbf{I}$ and $\lambda = \dfrac{\sigma_e^2}{\sigma_a^2} = \dfrac{1-h^2}{h^2}$.

The linear model is mixed in both the statistical sense (Henderson, 1984), as it contains fixed and random effects, and the genetic sense (Morton and MacLean, 1974), as it contains a single locus and a polygenic effect. Strictly additive gene

action of the polygenes is assumed but dominance is allowed for at the major locus. In order to keep the model simple, it is further assumed that the variance components σ_a^2 and σ_e^2 are known. This assumption implies that the genetic variance caused by polygenes is known but not the genetic variation caused by the segregating major allele, which is determined by the major genotype effects and frequencies. This critical assumption has to be kept in mind when discussing the simulation results.

Likelihood function

The likelihood for mixed model [1] was first discussed by Elston and Stewart (1971). The likelihood can be written as:

$$L(\mathbf{y}) = \sum_{\mathbf{T}} \mathrm{f}(\mathbf{y}|\mathbf{T},\mathbf{b},\mathbf{g},\lambda,\sigma_e^2) \cdot \Pr(\mathbf{T}|p), \qquad [2]$$

where

$$\mathrm{f}(\mathbf{y}|\mathbf{T},\mathbf{b},\mathbf{g},\lambda,\sigma_e^2) = \mathrm{c}_1 \cdot \mathrm{e}^{-0.5\cdot(\mathbf{y}-\mathbf{Xb}-\mathbf{ZTg})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{ZTg})\cdot\sigma_e^{-2}}$$

is a normal density and $\Pr(\mathbf{T}|p)$ is the probability of $\mathbf{T}$ given the allele frequency p and the pedigree information. Because variance components are assumed to be known, $\mathrm{c}_1 = (2\pi)^{-0.5n_0} \cdot |\mathbf{V} \cdot \sigma_e^2|^{-0.5}$, with n_0 as the number of observations, is a constant. Following Elston and Stewart (1971), $\Pr(\mathbf{T}|p)$ can be computed as a product of probabilities:

$$\Pr(\mathbf{T}|p) = \prod_{i}^{N} \Pr(t_i|t_s,t_d)$$

where N is the total number of animals in the population and $\Pr(t_i|t_s,t_d)$ is the probability of animal i having genotype indicated by t_i, the ith row of $\mathbf{T}$, given the genotypes of its parents s and d, and is assumed to be known. Elston and Stewart (1971) give $\Pr(t_i|t_s,t_d)$ for autosomal and sex-linked loci. When the parents are unknown $\Pr(t_i|t_s,t_d)$ is replaced by the frequency of the genotype t_i in the base population. Known major locus genotypes can be accomodated by setting $\Pr(t_i|t_s,t_d)$ to zero whenever t_i conflicts with the known genotype of animal i. With the base population (animals with unknown parents) in Hardy-Weinberg equilibrium, $\Pr(\mathbf{T}|p)$ can be written as:

$$\Pr(\mathbf{T}|p) = p^{2\cdot n_1} \cdot (2p(1-p))^{n_2} \cdot (1-p)^{2\cdot n_3} \cdot \prod_{i=n_b+1}^{N} \Pr(t_i|t_s,t_d),$$

where n_1, n_2 and n_3 are the number of base animals of genotype AA, Aa and aa, respectively, and $n_b = n_1 + n_2 + n_3$ is the total number of base animals.

With 3 possible genotypes the sum in [2] is over 3^N elements. For 20 animals the sum is already over 3.5×10^9 possible incidence matrices $\mathbf{T}$. Whenever $\mathbf{T}$ conflicts with the pedigree information $\Pr(\mathbf{T}|p)$ is zero. Therefore, depending on the pedigree structure, a large number of the elements to sum are zero, but there remains a considerable number of non-zero elements.

As pointed out by Elston and Stewart (1971) the 3 likelihoods conditional on an animal's genotype t_i are proportional to the probabilities of animal i having 1 of the 3 possible genotypes. The conditional likelihoods can be obtained by skipping animal i in the summation over all possible incidence matrices $\mathbf{T}$.

Maximum likelihood estimation

In order to maximize $L(\mathbf{y})$, we need the first derivatives with respect to $\mathbf{b}$, $\mathbf{g}$ and p:

$$\frac{\partial L(\mathbf{y})}{\partial \mathbf{b},\mathbf{g}} = \sum_{\mathbf{T}} c_1 \cdot e^{-0.5(\mathbf{y}-\mathbf{Xb}-\mathbf{ZTg})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{ZTg})\sigma_e^{-2}} \cdot [\mathbf{XZT}]'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{ZTg})\sigma_e^{-2} \cdot \Pr(\mathbf{T}|p)$$

$$\frac{\partial L(\mathbf{y})}{\partial p} = \sum_{\mathbf{T}} c_1 \cdot e^{-0.5(\mathbf{y}-\mathbf{Xb}-\mathbf{ZTg})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{ZTg})\sigma_e^{-2}} \cdot \Pr(\mathbf{T}|p)[(2n_1+n_2)\cdot p^{-1} - (n_2+2n_3)\cdot(1-p)^{-1}]$$

The probability of $\mathbf{T}$ given the data and the parameters of the model will be denoted $w_{\mathbf{T}}$ and can be computed as

$$w_{\mathbf{T}} = c_2 \cdot e^{-0.5\cdot(\mathbf{y}-\mathbf{Xb}-\mathbf{ZTg})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{ZTg})\cdot\sigma_e^{-2}} \cdot \Pr(\mathbf{T}|p),$$

where c_2 is the product of c_1 and a scaling factor such that $\sum_{\mathbf{T}} w_{\mathbf{T}} = 1$. Note that without scaling this sum is equal to the likelihood $L(\mathbf{y})$. After setting to zero and rearranging we get the 2 following equations:

$$\sum_{\mathbf{T}} w_{\mathbf{T}} \cdot \begin{bmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{V}^{-1}\mathbf{ZT} \\ \mathbf{T}'\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{T}'\mathbf{Z}'\mathbf{V}^{-1}\mathbf{ZT} \end{bmatrix} \cdot \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \sum_{\mathbf{T}} w_{\mathbf{T}} \cdot \begin{bmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{T}'\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y} \end{bmatrix}, \text{ and} \quad [3]$$

$$\sum_{\mathbf{T}} w_{\mathbf{T}}[(2n_1+n_2)\cdot\hat{p}^{-1} - (n_2+2n_3)\cdot(1-\hat{p})^{-1}] = 0.$$

Solving for $\hat{p}$ in the last equation leads to:

$$\hat{p} = \frac{1}{2n_b}\sum_{\mathbf{T}} w_{\mathbf{T}} \cdot (2n_1+n_2)$$

This equation can be rewritten by replacing $2n_1+n_2$ by $\mathbf{v}_b' \cdot \mathbf{T} \cdot [2\ 1\ 0]'$, with $\mathbf{v}_b'$ a row vector of length N with ones for base animals and zeros for the other animals.

$$\hat{p} = \frac{1}{2n_b} \cdot \mathbf{v}_b' \cdot \left(\sum_{\mathbf{T}} w_{\mathbf{T}} \cdot \mathbf{T}\right) \cdot [2\ 1\ 0]' \quad [4]$$

Because $w_{\mathbf{T}}$ depends on $\mathbf{b}$, $\mathbf{g}$ and p, equations [3] and [4] have to be solved iteratively. Let $w_{\mathbf{T}}^r$ be $w_{\mathbf{T}}$ with solutions for $\mathbf{b}$, $\mathbf{g}$ and p after round r replacing the

true values and $\mathbf{Q}^r = \sum_{\mathrm{T}} w_{\mathrm{T}}^r \mathbf{T}$. Note that the ikth element of $\mathbf{Q}^r$ at convergence is an estimate of the probability that animal i is of genotype k given the data and the estimates for the fixed effects $\widehat{\mathbf{b}}$, the major locus effects $\widehat{\mathbf{g}}$ and the allele frequency $\widehat{p}$. As mentioned above, the same estimate can be obtained by calculating likelihoods conditional on an animal's 3 genotypes. Using these definitions, equations [3] and [4] can be written as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathrm{X}'\mathrm{V}^{-1}\mathrm{ZQ}^{\mathrm{r}} \\ \mathbf{Q}^{r'}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X} & \sum_{\mathrm{T}} \mathrm{w}_{\mathrm{T}}^{\mathrm{r}} \cdot \mathrm{T}'\mathrm{Z}'\mathrm{V}^{-1}\mathrm{ZT} \end{bmatrix} \cdot \begin{bmatrix} \widehat{\mathbf{b}}^{r+1} \\ \widehat{\mathbf{g}}^{r+1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{Q}^{r'}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y} \end{bmatrix}, \text{ and} \quad [5]$$

$$\widehat{p}^{r+1} = \frac{1}{2n_{\mathrm{b}}} \cdot \mathbf{v}_{\mathrm{b}}' \cdot \mathbf{Q}^r \cdot [2\ 1\ 0]' \quad [6]$$

The solutions for $\widehat{\mathbf{b}}^r$, $\widehat{\mathbf{g}}^r$ and $\widehat{p}^r$ converge to maximum likelihood (ML) estimates. Local maxima in $L(\mathbf{y})$ could pose a problem and will be discussed later. Hoeschele (1988) estimated the allele frequency from the genotype probabilities of all animals with records whereas [6] considers only base animals, which is in agreement with Ott (1979). Because genotype probabilities of base animals take information from their descendants into account, all information on the allele frequency in the base populations is properly used by [6].

Animal breeders are not only interested in estimating major locus effects $\mathbf{g}$ and allele frequency p but also in predicting polygenic breeding values $\mathbf{a}$. This is usually done by regressing phenotypic observations corrected for fixed effects:

$$\widehat{a} = \lambda^{-1}\mathbf{A}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}} - \mathbf{Z}\widehat{\mathbf{Q}}\widehat{\mathbf{g}}),$$

where $\widehat{\mathbf{Q}}$ is $\mathbf{Q}^r$ at convergence. Using $\mathbf{V}^{-1} = [\mathbf{Z}\mathbf{A}\mathbf{Z}'\lambda^{-1} + \mathbf{I}]^{-1} = \mathbf{I} - \mathbf{Z}\mathbf{M}\mathbf{Z}'$, where $\mathbf{M} = [\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\lambda]^{-1}$ (Henderson, 1984), $\widehat{\mathbf{a}}$ can also be computed as:

$$\widehat{\mathbf{a}} = \mathbf{M}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}} - \mathbf{Z}\widehat{\mathbf{Q}}\widehat{\mathbf{g}}) \quad [7]$$

The same solutions for $\widehat{\mathbf{b}}$, $\widehat{\mathbf{g}}$ and $\widehat{\mathbf{a}}$ are obtained by iterating on the following equations together with [6] instead of using [5], [6] and [7]:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\mathbf{Q}^r & \mathbf{X}'\mathbf{Z} \\ \mathbf{Q}^{r'}\mathbf{Z}'\mathbf{X} & \mathbf{B}^r & \mathbf{Q}^{r'}\mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}\mathbf{Q}^r & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}}^{r+1} \\ \widehat{\mathbf{g}}^{r+1} \\ \widehat{\mathbf{a}}^{r+1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Q}^{r'}\mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \quad [8]$$

where $\mathbf{B}^r = \sum_{\mathrm{T}} w_{\mathrm{T}}^r \mathbf{T}'\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{T} + \mathbf{Q}^{r'}\mathbf{Z}'\mathbf{Z}\mathbf{M}\mathbf{Z}'\mathbf{Z}\mathbf{Q}^r$

$$= \sum_{\mathrm{T}} w_{\mathrm{T}}^r \mathbf{T}'\mathbf{Z}'\mathbf{Z}\mathbf{T} - \sum_{\mathrm{T}} w_{\mathrm{T}}^r \mathbf{T}'\mathbf{Z}'\mathbf{Z}\mathbf{M}\mathbf{Z}'\mathbf{Z}\mathbf{T} + \mathbf{Q}^{r'}\mathbf{Z}'\mathbf{Z}\mathbf{M}\mathbf{Z}'\mathbf{Z}\mathbf{Q}^r$$

Note that $\sum_{\mathrm{T}} w_{\mathrm{T}}^r \mathbf{T}'\mathbf{Z}'\mathbf{Z}\mathbf{T} = \mathrm{diag}\{\mathbf{v}_0' \cdot \mathbf{q}_k^r\} = \mathbf{D}^r$, where $\mathbf{v}_0'$ is a row vector containing the diagonal elements of $\mathbf{Z}'\mathbf{Z}$ and $\mathbf{q}_k^r$ the kth column of $\mathbf{Q}^r$. The

difficulty with this approach is that it is not feasible to compute $\mathbf{Q}^r$ and $\sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot \mathbf{T}'\mathbf{Z}'\mathbf{Z}\mathbf{M}\mathbf{Z}'\mathbf{Z}\mathbf{T}$ for large populations.

Approximations

Above $\mathbf{Q}^r$ was defined as:

$$\mathbf{Q}^r = \sum_{\mathbf{T}} w_{\mathbf{T}}^r \mathbf{T} = \sum_{\mathbf{T}} c_2 \cdot e^{-0.5\cdot(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^r-\mathbf{Z}\mathbf{T}\hat{\mathbf{g}}^r)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^r-\mathbf{Z}\mathbf{T}\hat{\mathbf{g}}^r)\cdot\sigma_e^{-2}} \cdot \Pr(\mathbf{T}|\hat{p}^r) \cdot \mathbf{T}$$

There are 2 problems associated with the computation of $\mathbf{Q}^r$. Firstly, the summation is over all possible incidence matrices $\mathbf{T}$ and, secondly, a quadratic form involving $\mathbf{V}^{-1}$ has to be computed for each element in this sum. It can be shown that the following is an equivalent expression not involving $\mathbf{V}^{-1}$:

$$\mathbf{Q}^r = \sum_{\mathbf{T}} c_2 \cdot e^{-0.5\cdot\hat{\mathbf{a}}_{\mathbf{T}}^{r\prime}\mathbf{A}^{-1}\hat{\mathbf{a}}_{\mathbf{T}}^r\cdot\sigma_a^{-2}} \cdot e^{-0.5\cdot(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^r-\mathbf{Z}\mathbf{T}\hat{\mathbf{g}}^r-\mathbf{Z}\hat{\mathbf{a}}_{\mathbf{T}}^r)'(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^r-\mathbf{Z}\mathbf{T}\hat{\mathbf{g}}^r-\mathbf{Z}\hat{\mathbf{a}}_{\mathbf{T}}^r)\cdot\sigma_e^{-2}} \cdot \Pr(\mathbf{T}|\hat{p}^r) \cdot \mathbf{T},$$

where $\hat{\mathbf{a}}_{\mathbf{T}}^r = \mathbf{M}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^r - \mathbf{Z}\mathbf{T}\hat{\mathbf{g}}^r)$ (Le Roy *et al*, 1989). Because $\hat{\mathbf{a}}_{\mathbf{T}}^r$ depends on $\mathbf{T}$, we would have to compute $\hat{\mathbf{a}}_{\mathbf{T}}^r$ for every possible $\mathbf{T}$, which is not feasible. In order to simplify the computations, we could replace $\hat{\mathbf{a}}_{\mathbf{T}}^r$ by $\hat{\mathbf{a}}^r$ which does not depend on $\mathbf{T}$. Note that $\hat{\mathbf{a}}^r = \sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot \hat{\mathbf{a}}_{\mathbf{T}}^r$. This approximation was also considered by Hoeschele (1988). The approximated $\mathbf{Q}^r$ is then:

$$\overset{\circ}{\mathbf{Q}}{}^r = \sum_{\mathbf{T}} c_2 \cdot e^{-0.5\cdot\hat{\mathbf{a}}^{r\prime}\mathbf{A}^{-1}\hat{\mathbf{a}}^r\cdot\sigma_a^{-2}} \cdot e^{-0.5\cdot(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^r-\mathbf{Z}\mathbf{T}\hat{\mathbf{g}}^r-\mathbf{Z}\hat{\mathbf{a}}^r)(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^r-\mathbf{Z}\mathbf{T}\hat{\mathbf{g}}^r-\mathbf{Z}\hat{\mathbf{a}}^r)\cdot\sigma_e^{-2}} \cdot \Pr(\mathbf{T}|\hat{p}^r) \cdot \mathbf{T} \qquad [9]$$

Instead of using a single estimate of the polygenic breeding value for each animal irrespective of its genotype, we could use 3 values for each animal depending on its genotype but independent of the genotypes of all the other animals. A similar approximation was considered by Elsen and Le Roy (1989) and Knott *et al* (1992a, 1992b) for a sire model and was found to be superior to [9]. We considered the following approximation:

$$\tilde{\mathbf{Q}}^r = \sum_{\mathbf{T}} c_2 \cdot e^{-0.5\cdot\hat{\mathbf{a}}^{r\prime}\mathbf{A}^{-1}\hat{\mathbf{a}}^r\sigma_a^{-2}} \cdot e^{-0.5\cdot(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^r-\mathbf{Z}\mathbf{T}\hat{\mathbf{g}}^r-\mathbf{Z}\tilde{\mathbf{a}}_{\mathbf{T}}^r)'(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^r-\mathbf{Z}\mathbf{T}\hat{\mathbf{g}}^r-\mathbf{Z}\tilde{\mathbf{a}}_{\mathbf{T}}^r\cdot\sigma_e^{-2}} \cdot \Pr(\mathbf{T}|\hat{p}^r) \cdot \mathbf{T} \qquad [10]$$

where $\tilde{\mathbf{a}}_{t_{ik}}^r$ the element of $\tilde{\mathbf{a}}_{\mathbf{T}}^r$ for animal i with genotype k is calculated as:

$$\tilde{\mathbf{a}}_{t_{ik}}^r = (\mathbf{y}_i - \mathbf{x}_i\hat{\mathbf{b}}^r - \mathbf{t}_{ik}\hat{\mathbf{g}}^r - \sum_{j\neq i}^{N} \mathbf{a}^{ij} \cdot \hat{\mathbf{a}}_j^r / c_{ii})$$

where $\mathbf{x}_i$ and $\mathbf{t}_{ik}$ are the ith rows of $\mathbf{X}$ and $\mathbf{ZT}$, a^{ij} is the ijth element of $\mathbf{A}^{-1}$, and c_{ii} is the diagonal element of the coefficient matrix in [8] pertaining to the ith animal equation.

The summation over all possible incidence matrices $\mathbf{T}$ in [9] or [10] can be avoided by using algorithms developed to estimate genotype probabilities. Here, the iterative algorithm of van Arendonk *et al* (1989) was applied. This procedure will be briefly described in the next section.

As with $\mathbf{Q}^r$ the difficulty with expression $\sum w_{\mathbf{T}}^r \cdot \mathbf{T}'\mathbf{Z}'\mathbf{ZMZ}'\mathbf{ZT}$ is two-fold; the sum is over all possible $\mathbf{T}$, and the computation of each element in that sum is expensive. Let m_{ij} be the ijth element of $\mathbf{Z}'\mathbf{ZMZ}'\mathbf{Z}$, and $t_{ik}(t_{jl})$ be the elements of $\mathbf{T}$ for animal $i(j)$ and genotype $k(l)$. Now, the klth element of $\sum w_{\mathbf{T}}^r \mathbf{T}'\mathbf{Z}'\mathbf{ZMZ}'\mathbf{ZT}$ can be calculated as:

$$\sum_{\mathbf{T}} w_{\mathbf{T}}^r \sum_i \Big(\sum_j t_{jl} \cdot m_{ij} \Big) \cdot t_{ik} = \sum_i \sum_j \Big(\sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot t_{ik} \cdot t_{jl} \Big) \cdot m_{ij}$$

Note that at convergence $\sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot t_{ik} \cdot t_{jl}$ is an estimate of the probability that animal i is of genotype k and animal j of genotype l, given the data. For independent animals this quantity is equal to $q_{ik}^r \cdot q_{jl}^r$ the product of the corresponding elements in $\mathbf{Q}^r$ and, therefore, the contributions of $\sum_{\mathbf{T}} w_{\mathbf{T}}^r \mathbf{T}'\mathbf{Z}'\mathbf{ZMZ}'\mathbf{ZT}$ and $\mathbf{Q}^{r'}\mathbf{Z}'\mathbf{ZMZ}'\mathbf{ZQ}^r$ to $\mathbf{B}^r$ cancel out. For dependent animals the contributions to the klth element of $\mathbf{B}^r$ are:

$$\sum_i \sum_j q_{ik}^r \cdot q_{jl}^r m_{ij} - \sum_i \sum_j \Big(\sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot t_{ik} \cdot t_{jl} \Big) \cdot m_{ij}$$
$$= \sum_i \sum_j \Big(q_{ik}^r \cdot q_{jl}^r - \Big(\sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot t_{ik} \cdot t_{jl} \Big) \Big) \cdot m_{ij}$$

Now if we neglect the dependencies between animals for the computation of $\sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot t_{ik} \cdot t_{jl}$ we get:

$$\overset{\circ}{\mathbf{B}}{}^r = \sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot \mathbf{T}'\mathbf{Z}'\mathbf{ZT} = \mathbf{D}^r \qquad [11]$$

and [8] becomes identical to the mixed model equations given by Hoeschele (1988).

Another way to approximate $\mathbf{B}^r$ is to assume that $\mathbf{A} = \mathbf{I}$. We then get:

$$\mathbf{ZMZ}' = \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{I}\lambda)^{-1}\mathbf{Z}' = (1 + \lambda)^{-1} \cdot \mathbf{I} = h^2 \cdot \mathbf{I},$$

and $\mathbf{B}^r$ simplifies to:

$$\widetilde{\mathbf{B}}^r = \sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot \mathbf{T}'\mathbf{Z}'\mathbf{ZT} - h^2 \cdot \sum_{\mathbf{T}} w_{\mathbf{T}}^r \cdot \mathbf{T}'\mathbf{Z}'\mathbf{ZT} + h^2 \cdot \mathbf{Q}^{r'}\mathbf{Z}'\mathbf{ZQ}^r$$
$$= (1 - h^2) \cdot \mathbf{D}^r + h^2 \cdot \mathbf{Q}^{r'}\mathbf{Z}'\mathbf{ZQ}^r \qquad [12]$$

Estimation of genotype probabilities

Van Arendonk *et al* (1989) developed an iterative algorithm to estimate genotype probabilities for discrete phenotypes. Kinghorn *et al* (1993) applied this algorithm to continuous traits. The comparison of this algorithm with non-iterative methods revealed some errors in the formulae given in the original paper (LLG Janss and JAM van Arendonk, 1991; C Stricker, 1992; personal communications). We applied a corrected version of this algorithm.

For each animal, genotype probabilities from 3 different sources of information are computed using approximation [9] or [10]. One round of iteration involves 3 steps. First genotype probabilities are computed using information from parents and collateral relatives proceeding from the oldest to the youngest animal. In the second step, genotype probabilities are calculated using information from the progeny proceeding from the youngest to the oldest animal. Finally, genotype probabilities using information from each individual performance are calculated and the 3 sources of information combined. The iteration process is stopped when the solutions for genotype probabilities reach a given convergence criterion.

The algorithm works for simpler pedigree structures as simulated in this study but does not allow for loops in the pedigree, also known as cycles (Lange and Elston, 1975). Loops in a pedigree occur through genetic paths (inbreeding loops), mating paths, or a combination of the 2 (marriage loops), *eg*, a sire mated to 2 genetically related dams. Both inbreeding and marriage loops are common in animal breeding data. A non-iterative algorithm for pedigrees without loops was recently proposed, which should be more efficient than the one used in this study (Fernando *et al*, 1993).

Method of Hoeschele (1988)

Hoeschele (1988) used a Bayesian approach to derive an iterative procedure to estimate genotype probabilities $\mathbf{Q}$, allele frequency p and major locus effects $\mathbf{g}$ for simple pedigree structures. The genotype probabilities were estimated by formulae that were developed for the specific pedigree structures considered using approximation [9]. In contrast to [6], Hoeschele (1988) estimated p from the genotype probabilities of all animals with records:

$$\widehat{p}^{r+1} = \frac{1}{2n_0} \cdot \mathbf{v}_0' \cdot \mathbf{Q}^r \cdot [2\ 1\ 0]' \qquad [13]$$

where n_0 is the number of animals with records and $\mathbf{v}_0'$ is a row vector with ones for animals with records and zeros otherwise. The equations that estimate the effects of model [1] are the same as [8] approximated with [11]. We applied this method in the simulation study using the iterative algorithm described above but with approximation [9] to estimate genotype probabilities instead of the formulae given by Hoeschele.

Method of Kinghorn et al *(1993)*

In least-squares analysis it is usually assumed that all independent variables are known without error. When independent variables are measured with some error,

the least-squares estimates are biased (see, for example, Johnston, 1984, p 428). Kinghorn *et al* (1993) treated the unknown incidence matrix $\mathbf{T}$ as the unknown true independent variable and the genotype probabilities $\mathbf{Q}$ as an estimate for $\mathbf{T}$ associated with some errors. Using $\mathbf{Q}$ instead of $\mathbf{T}$ in the model leads to biased estimates of $\widehat{\mathbf{g}}^*$. Kinghorn *et al* (1993) derived a correction matrix $\mathbf{W}$, such that $\widehat{\mathbf{g}} = \mathbf{W}^{-1}\widehat{\mathbf{g}}^*$. Given certain assumptions, they showed that $\mathbf{W} = \mathbf{V}_q^{-1}\mathbf{V}_t$, where $\mathbf{V}_t$ is a 3×3 covariance matrix of elements in the 3 columns of $\mathbf{T}$ and $\mathbf{V}_q$ is the corresponding covariance matrix of elements in the 3 columns of $\mathbf{Q}$. Because (co)variances in $\mathbf{V}_q$ are generally smaller than (co)variances in $\mathbf{V}_t$, major locus effects are overestimated in absolute terms when using $\mathbf{Q}$ instead of $\mathbf{T}$. The (co)variances in $\mathbf{V}_q$ were calculated from the actual solutions for estimates of genotype probabilities of all animals with records. Covariances in $\mathbf{V}_t$ were computed as:

$$\text{Cov}(t_k,t_k) = \overline{q}_{.k} \cdot (1 - \overline{q}_{.k}), \text{ and } \text{Cov}(t_k,t_l) = -\overline{q}_{.k} \cdot \overline{q}_{.l} \quad \text{for} \quad k \neq l, \qquad [14]$$

where $\overline{q}_{.k}$ is the average genotype probability for genotype k of all animals with records and can be regarded as an estimate of the frequency of that genotype in the population. Genotype probabilities were estimated with the algorithm of van Arendonk *et al* (1989). This algorithm requires the allele frequency p as an input parameter. Kinghorn *et al* (1993) kept the initial value for p constant over all iterations, *ie* regarded the initial p as the true value. But if p was known, $\text{Cov}(t_k,t_1)$ could also be derived from the expected frequencies of the 3 genotypes. In our implementation $\text{Cov}(t_k,t_l)$ was computed with [14] and the allele frequency p was estimated with [13], which is a natural deduction from [14].

The linear model can be written in matrix notation as:

$$\mathbf{y} = \mathbf{X}\mathbf{b}^* + \mathbf{Z}\mathbf{Q}\mathbf{W}\mathbf{g} + \mathbf{Z}\mathbf{a}^* + \mathbf{e}^*$$

Kinghorn *et al* (1993) assumed that $\text{Var}(\mathbf{a}^*) = \text{Var}(\mathbf{a}) = \mathbf{A} \cdot \sigma_a^2$ and $\text{Var}(\mathbf{e}^*) = \text{Var}(\mathbf{e}) = \mathbf{I} \cdot \sigma_e^2$. The matrices $\mathbf{Q}$ and $\mathbf{W}$ are not known and have to be estimated from the data as described above. Therefore, the following system of equations has to be solved iteratively:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\mathbf{Q}^r\mathbf{W}^r & \mathbf{X}'\mathbf{Z} & \\ \mathbf{W}^{r'}\mathbf{Q}^{r'}\mathbf{Z}'\mathbf{X} & \mathbf{W}^{r'}\mathbf{Q}^{r'}\mathbf{Z}'\mathbf{Z}\mathbf{Q}^r\mathbf{W}^r & \mathbf{W}^{r'}\mathbf{Q}^{r'}\mathbf{Z}'\mathbf{Z} & \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}\mathbf{Q}^r\mathbf{W}^r & \mathbf{Z}'\mathbf{Z} & +\mathbf{A}^{-1}\lambda \end{bmatrix} \cdot \begin{bmatrix} \widehat{\mathbf{b}}^{*r+1} \\ \widehat{\mathbf{g}}^{r+1} \\ \mathbf{a}^{*r+1} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}^{r'}\mathbf{Q}^{r'}\mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \qquad [15]$$

Estimates for $\mathbf{g}$ should be unbiased but estimates for $\mathbf{b}$ and $\mathbf{a}$ are still biased. We attempted to correct for the bias in $\mathbf{b}$ by adding $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{Q}(\mathbf{W} - \mathbf{I})\widehat{\mathbf{g}}^{r+1}$, the expected difference between $\widehat{\mathbf{b}}^{r+1}$ and $\widehat{\mathbf{b}}^{*r+1}$ under the assumptions $\mathbf{E}(\mathbf{T}) = \mathbf{E}(\mathbf{Q})$, $\mathbf{E}(\mathbf{a} - \mathbf{a}^*) = 0$, and $\mathbf{E}(\mathbf{e} - \mathbf{e}^*) = 0$, to the current solution $\widehat{\mathbf{b}}^{*r+1}$.

Simulation

The methods of Hoeschele (1988) and Kinghorn *et al* (1993) were compared with the method developed in this study applying approximations [10] and [12] using stochastic computer simulation. Phenotypic observations were generated by using the following mixed model:

$$y_{ijk} = hys_i + g_j + a_{ijk} + e_{ijk},$$

where hys_i is the fixed effect of herd × year × sex i, g_j is the fixed effect of major locus genotype j, a_{ijk} is the polygenic breeding value and e_{ijk} is the random residual effect. The effects in the model were sampled as follows: $\{hys_i\} \sim N(0,\mathbf{I}\sigma_h^2)$ $\{a_{ijk}\} \sim N(0,\mathbf{A}\sigma_a^2)$ and $\{e_{ijk}\} \sim N(0,\mathbf{I}\sigma_e^2)$. Major locus genotypes were simulated with 2 segregating alleles. Genotypes of base animals were generated by sampling 2 alleles from a uniform distribution between 0.0 and 1.0 with threshold p, the frequency of allele A. Genotypes of progeny were determined according to mendelian segregation. The effect of genotype 3 was set to zero as there is a dependency between fixed herd × year × sex and major locus effects.

Three different sets of parameters were used (table I). Only additive effects of the major locus were considered, although all of the methods compared allow for dominance. In the first set of parameters, 50% of the phenotypic variance (variance due to major locus + polygenic variance + residual variance) is due to genetic effects, 75% of the genetic variance is due to the major locus, and 25% is due to the polygenes. The frequency of allele A with major effect is 25% in the base population, which results in an allele substitution effect α of 1.0, *ie* genotype effects of 2.0 (AA), 1.0 (Aa) and 0 (aa). In parameter set 2, the allele frequency p is 0.5, but the genotype effects and all the other parameters are the same as in set 1. Thus the variance due to the major locus is increased from 0.375 to 0.5, and the phenotypic variance changes from 1.0 to 1.125. In parameter set 3, the allele frequency p is 0.25 and 50% of the phenotypic variance is due to genetic effects, as in parameter set 1, but the proportion of genetic variance due to the polygenes is increased from 25 to 40%, which results in an allele substitution effect α of $\sqrt{0.8}$.

Table I. True values for the 3 parameter sets used in the simulation study.

Parameter	*1*	*2*	*3*
g_1	2.000	2.000	1.789
g_2	1.000	1.000	0.894
p	0.250	0.500	0.250
σ_a^2	0.125	0.125	0.200
σ_e^2	0.500	0.500	0.500
σ_h^2	1.000	1.000	1.000

Because the algorithm to estimate genotype probabilities used in this study does not allow for complex pedigrees, the structure of the simulated population is very

simple. In each of 10 herds, 20 base dams each had a record in year 1. A group of 20 base sires each with their own record in a common herd × year (*eg* test station) was mated with these base dams. Each sire was randomly mated with 1 dam in each herd. Each mating produced 5 progeny in year 2. The sex of each progeny was determined by sampling from a uniform distribution between 0.0 and 1.0 with threshold 0.5. The population size was 1 220, made up of 220 base animals and 1 000 progeny.

In each of the alternatives, the same sequence of random numbers was used. Therefore, identical data sets were analysed with each of the 3 methods considered. Each alternative was replicated 25 times.

With each of the 3 methods, final solutions are obtained by repeatedly computing genotype probabilities and solving a system of equations to get new solutions for major genotype effects and polygenic breeding values. A stopping criterion of the form:

$$\sqrt{\frac{(\hat{\mathbf{g}}^r - \hat{\mathbf{g}}^{r-1})'(\hat{\mathbf{g}}^r - \hat{\mathbf{g}}^{r-1})}{\hat{\mathbf{g}}^{r'}\hat{\mathbf{g}}^r}} < 10^{-4}$$

was used for major genotype effects $\mathbf{g}$ and the allele frequency p.

RESULTS

When the genotypes of all animals with records are known, the estimates for major locus effects $\mathbf{g}$ are identical for all 3 methods considered (table II). Estimates for the allele frequency p, however, differed slightly. Using formula [13] (Hoeschele, 1988; Kinghorn *et al*, 1993) the standard deviations (SD) of estimated p were larger than estimates by [6]. The estimates for $\mathbf{g}$ and p agree well with the true values. Estimates of $\mathbf{g}$ across parameter sets are consistently slightly larger than the true values, which can be explained by sampling effects and the fact that for each of the 25 replicates, data for the 3 parameter sets were generated with the same set of random numbers. As expected from the heritabilities, the correlations between true and predicted breeding values were the same for parameter sets 1 and 2 and slightly higher for parameter set 3. The correlations between predicted breeding values and estimated major locus effects were close to zero, showing that the 2 effects were well separated in all cases.

Table III shows the simulation results for the 3 parameter sets using all 3 procedures when major locus genotypes were unknown. For parameter sets 1 and 2, estimates of major locus effects $\mathbf{g}$ were close to the true values or slightly underestimated with approximated maximum likelihood (AML), underestimated by about 20% with the method of Hoeschele (1988) and overstimated by 25 to 30% with the method of Kinghorn *et al* (1993). For parameter set 3, estimates of major locus effects $\mathbf{g}$ were zero for 2 replicates using AML and for 21 replicates using the method of Hoeschele (1988). Non-zero estimates of $\mathbf{g}$ were biased upwards by 14% with AML and by 47% with the method of Kinghorn *et al* (1993). Both AML and the method of Hoeschele (1988) showed a large variability of the non-zero estimates of major locus effects for parameter set 3. When the true allele frequency was 0.25 the allele frequency p was substantially underestimated with

Table II. Estimates of major locus effects **g** and allele frequency p, and accuracies of predicted polygenic breeding values and major locus effects with known major locus genotypes for the 3 parameter sets (footnote 1).

	1		2		3	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
$\widehat{\mathbf{g}}_1$	2.069	0.068	2.057	0.061	1.855	0.073
$\widehat{\mathbf{g}}_2$	1.027	0.047	1.015	0.047	0.921	0.051
$\widehat{p}^2_{[6]}$	0.251	0.018	0.499	0.020	0.251	0.018
$\widehat{p}^2_{[13]}$	0.253	0.025	0.502	0.030	0.253	0.025
$r_{\mathbf{a},\hat{\mathbf{a}}}$	0.626	0.044	0.625	0.044	0.687	0.034
$r_{\mathbf{T_g},\mathbf{T}_{\hat{\mathbf{g}}}}$	1.000	0.001	1.000	0.000	0.999	0.001
$r_{\hat{\mathbf{a}},\mathbf{T}_{\hat{\mathbf{g}}}}$	0.009	0.092	−0.001	0.087	0.009	0.086

[1] Mean and standard deviation over 25 replicates; [2] using formulae [6] or [13].

AML, but estimated quite well with the other 2 methods. Correlations between true and predicted breeding values were similar for AML and the method of Hoeschele (1988), but zero for the method of Kinghorn *et al* (1993). For parameter sets 1 and 2, the correlations between true ($\mathbf{Tg}$) and estimated ($\widehat{\mathbf{Q}}\widehat{\mathbf{g}}$) major locus effects were similar for all 3 methods. When major locus effects were smaller (parameter set 3) these correlations were largest with the method of Kinghorn *et al* (1993). Predicted breeding values were positively correlated to estimated major locus effects $\widehat{\mathbf{Q}}\widehat{\mathbf{g}}$ with AML and to a larger extent with the method of Hoeschele (1988). Using the method of Kinghorn *et al* (1993) these correlations were strongly negative.

Because poor estimation of p also affects all the other estimates, additional simulations were done with the allele frequency fixed at the true (expected) value. Results are reported in table IV for AML and the method of Hoeschele (1988) for parameter sets 1 and 3. All other results were close to those of table III and are therefore not shown. Major locus effects **g** were underestimated less with AML and the correlations were similar for both methods. For parameter set 3, the number of replicates with estimates of zero for major locus effects was again much larger with the method of Hoeschele (1988).

Table V compares the 3 methods for the case where all sires and 50% of the dams are genotyped at the major locus. There was still a tendency for AML to underestimate the allele frequency p when the true frequency was 0.25. The method of Hoeschele (1988) underestimated major locus effects considerably more than AML (9 to 31% *versus* 1 to 11%), whereas these effects were overestimated by 22 to 43% with the method of Kinghorn *et al* (1993). The accuracies of predicted breeding values were again similar for AML and the method of Hoeschele (1988) but much lower for the method of Kinghorn *et al* (1993). The accuracies of estimated genetic values at the major locus were similar for all 3 methods with a tendency of lower accuracies for the method of Kinghorn *et al* (1993). When all the sires but none of the dams were genotyped the results, which are not reported here, were

Table III. Estimates of major locus effects **g** and allele frequency p, and accuracies of predicted polygenic breeding values and estimated major locus effects with unknown major locus genotypes using maximum likelihood with approximations [10] and [12] (AML) and the methods of Hoeschele (1988) and Kinghorn *et al* (1993) (footnote 1).

	AML		*Hoeschele*		*Kinghorn*	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Parameter set 1						
$\widehat{\mathbf{g}}_1$	2.058	0.237	1.659	0.189	2.607	0.132
$\widehat{\mathbf{g}}_2$	1.067	0.095	0.744	0.172	1.302	0.072
$\widehat{p}$	0.115	0.033	0.235	0.044	0.246	0.033
$r_{\mathbf{a},\hat{\mathbf{a}}}$	0.394	0.062	0.405	0.078	0.023	0.125
$r_{\mathbf{Tg},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.684	0.061	0.720	0.058	0.701	0.059
$r_{\hat{\mathbf{a}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.385	0.128	0.632	0.101	−0.627	0.028
Parameter set 2						
$\widehat{\mathbf{g}}_1$	1.836	0.085	1.628	0.089	2.495	0.074
$\widehat{\mathbf{g}}_2$	0.894	0.086	0.779	0.101	1.227	0.068
$\widehat{p}$	0.497	0.103	0.498	0.048	0.496	0.043
$r_{\mathbf{a},\hat{\mathbf{a}}}$	0.377	0.076	0.375	0.068	−0.060	0.144
$r_{\mathbf{T}\hat{\mathbf{g}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.752	0.035	0.752	0.035	0.729	0.040
$r_{\hat{\mathbf{a}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.614	0.085	0.711	0.066	−0.647	0.031
Parameter set 3[2]						
$\widehat{\mathbf{g}}_1$	2.042	0.686	1.779	0.914	2.664	0.114
$\widehat{\mathbf{g}}_2$	1.019	0.264	0.272	0.257	1.300	0.074
$\widehat{p}$	0.041	0.024	0.181	0.101	0.253	0.027
$r_{\mathbf{a},\hat{\mathbf{a}}}$	0.468	0.060	0.457	0.084	0.001	0.126
$r_{\mathbf{Tg},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.455	0.147	0.486	0.134	0.609	0.078
$r_{\hat{\mathbf{a}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.236	0.130	0.420	0.178	−0.649	0.029

[1] Mean and standard deviation over 25 replicates, starting values = true values; [2] included are only 23 (AML) and 4 (Hoeschele) replicates with non-zero estimates for **g**.

intermediate between the 2 cases of no animals and all sires plus 50% of the dams genotyped.

So far, final solutions have been reported for iterations where starting values were equal to true (expected) values. Table VI shows the number of replicates that converged to the same solutions using different starting values. Low starting values were half the true values and high starting values were 1.5 times the true values of major locus effects **g** and allele frequency p. When major locus genotypes were not known, none to a few replicates converged to a single set of solutions with all 3 different starting values. For the method of Hoeschele (1988) with parameter set 3, most of the replicates that converged to the same solutions converged to an estimate of zero for major locus effects **g**. For AML and the method of Hoeschele (1988), all replicates with 1 exception converged to 1 set of solutions when genotypes of all

Table IV. Estimates of major locus effects **g**, and accuracies of predicted polygenic breeding values and estimated major locus effects with known allele frequency and unknown major locus genotypes using maximum likelihood with approximations [10] and [12] (AML) and the method of Hoeschele (1988) (footnote 1).

	AML		*Hoeschele*	
	Mean	*SD*	*Mean*	*SD*
Parameter set 1				
$\hat{\mathbf{g}}_1$	1.866	0.157	1.645	0.155
$\hat{\mathbf{g}}_2$	0.879	0.096	0.717	0.138
$r_{\mathbf{a},\hat{\mathbf{a}}}$	0.412	0.075	0.407	0.080
$r_{\mathbf{Tg},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.736	0.049	0.724	0.060
$r_{\hat{\mathbf{a}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.596	0.084	0.657	0.085
Parameter set 3[2]				
$\hat{\mathbf{g}}_1$	1.536	0.233	1.386	0.153
$\hat{\mathbf{g}}_2$	0.467	0.283	0.219	0.024
$r_{\mathbf{a},\hat{\mathbf{a}}}$	0.499	0.075	0.519	0.038
$r_{\mathbf{Tg},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.580	0.100	0.590	0.079
$r_{\hat{\mathbf{a}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.615	0.124	0.534	0.142

[1] Mean and standard deviation over 25 replicates, starting values = true values; [2] included are only 17 (AML) and 2 (Hoeschele) replicates of non-zero estimates for **g**.

the sires (but none of the dams) were known. The largest number of replicates with all 3 solutions different was found with the method of Kinghorn *et al* (1993).

DISCUSSION

The method proposed here (AML) generally slightly underestimates major locus effects **g** and seriously underestimates allele frequency p when the true frequency is 0.25. The underestimation of p leads to increased estimates of **g**, although not to the extent that the variance explained by the major locus stays constant (tables III and IV). This variance is higher when the allele frequency is fixed at the true value. The allele frequency was still considerably underestimated for parameter set 1 when the population size was 10 times larger than considered here (results not shown). The allele frequency was estimated by [6], which was derived by maximizing the likelihood of the data, whereas the other 2 methods used [13]. Additional simulation runs with parameter sets 1 and 3 and approximations [9] and [11] together with [6] showed considerably lower estimates of p and higher estimates of **g** than results for the same 2 approximations applied together with [13], the method of Hoeschele (1988) (results not shown). There seems to be a problem in applying [6] together with approximations [10] and [12] or, to a lesser extent, with [9] and [11]. Nevertheless [6] is the correct equation for the estimation of the allele frequency by maximum likelihood.

Table V. Estimates for major locus effects $\mathbf{g}$ and allele frequency p, and accuracies of predicted polygenic breeding values and estimated major locus effects, with all sires and 50% of dams genotyped for the major locus using maximum likelihood with approximations [10] and [12] (AML) and the methods of Hoeschele (1988) and Kinghorn *et al* (1993) (footnote 1).

	AML		*Hoeschele*		*Kinghorn*	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Parameter set 1						
$\widehat{\mathbf{g}}_1$	1.981	0.108	1.823	0.117	2.498	0.104
$\widehat{\mathbf{g}}_2$	0.964	0.065	0.874	0.069	1.278	0.054
$\widehat{p}$	0.231	0.023	0.245	0.031	0.249	0.029
$r_{\mathbf{a},\widehat{\mathbf{a}}}$	0.560	0.054	0.551	0.057	0.329	0.082
$r_{\mathbf{Tg},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.821	0.027	0.820	0.027	0.811	0.031
$r_{\hat{\mathbf{a}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.329	0.090	0.410	0.084	−0.515	0.051
Parameter set 2						
$\widehat{\mathbf{g}}_1$	1.939	0.078	1.788	0.082	2.462	0.067
$\widehat{\mathbf{g}}_2$	0.954	0.067	0.877	0.072	1.218	0.057
$\widehat{p}$	0.497	0.022	0.502	0.036	0.500	0.033
$r_{\mathbf{a},\hat{\mathbf{a}}}$	0.538	0.059	0.528	0.060	0.258	0.084
$r_{\mathbf{Tg},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.827	0.017	0.825	0.017	0.819	0.018
$r_{\hat{\mathbf{a}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.398	0.103	0.476	0.092	−0.547	0.039
Parameter set 3						
$\widehat{\mathbf{g}}_1$	1.696	0.138	1.400	0.161	2.468	0.109
$\widehat{\mathbf{g}}_2$	0.796	0.077	0.620	0.079	1.276	0.061
$\widehat{p}$	0.226	0.023	0.244	0.032	0.250	0.029
$r_{\mathbf{a},\hat{\mathbf{a}}}$	0.621	0.044	0.600	0.050	0.378	0.075
$r_{\mathbf{Tg},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.784	0.033	0.780	0.035	0.773	0.036
$r_{\hat{\mathbf{a}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$	0.353	0.085	0.474	0.076	−0.516	0.052

[1] Mean and standard deviation over 25 replicates, starting values = true values.

The method of Hoeschele (1988) consistently underestimated major locus effects $\mathbf{g}$ which is in agreement with the simulation results of the same author. For smaller allele effects (parameter set 3), although still quite large, most of the estimates of $\mathbf{g}$ were zero, indicating that the genotype effects have to be large in order to be recognized. The same is true for AML, but to a lesser extent. There was a tendency for the accuracies of predicted polygenic breeding values ($\widehat{\mathbf{a}}$) and estimated major locus effects ($\widehat{\mathbf{O}}\widehat{\mathbf{g}}$) to be slightly higher with AML than with the method of Hoeschele (1988). In an unselected population as simulated here the expected correlation between true polygenic and major locus effects is zero. The correlations between the 2 estimates were positive for both methods but in almost all cases they were lower with AML. This indicates that the 2 estimates are less confounded with AML. With selection a negative correlation between the true effects will build up

Table VI. Number of replicates out of 25 that converged to the same solutions (max. abs. difference = 1%) of major locus effects **g** and allele frequency p for different starting values[1] using maximum likelihood with approximations [10] and [12] (AML) and the methods of Hoeschele (1988) and Kinghorn *et al* (1993) (allele frequency p estimated).

Genotyped	*AML*				*Hoeschele*				*Kinghorn*			
	l=t=h	*l=t*	*t=h*	*l=h*	*l=t=h*	*l=t*	*t=h*	*l=h*	*l=t=h*	*l=t*	*t=h*	*l=h*
Parameter set 1												
None	2	7	5	2	7	4	4	0	0	0	1	0
Sires	25	0	0	0	25	0	0	0	12	3	9	0
Parameter set 2												
None	0	1	0	0	0	0	0	0	0	0	0	0
Sires	24	1	0	0	25	0	0	0	18	2	3	0
Parameter set 3												
None	5^2	7	2	2	23^3	1	0	1^4	0	1	0	0
Sires	25	0	0	0	25	0	0	0	2	0	12	0

[1] Starting values: true (t) (= true simulation parameters), low (l) (= 0.5 × true) and high (h)
(= 1.5 × true); included are 2^2, 21^3 and 1^4 replicates with estimates of zero for **g**.

(gametic disequilibrium) which will make separation of the 2 effects more difficult. For AML and the method of Hoeschele (1988), the mean correlations $r_{\hat{\mathbf{a}},\mathbf{a}}$ were lower and $r_{\hat{\mathbf{a}},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$ were higher when the allele frequency was 0.5 (parameter set 2) than when the same allele had a frequency of 0.25 (parameter set 1) (tables III and V). Although the proportion of variance explained by the major locus is higher with parameter set 2 it seems to be more difficult to separate polygenic and major locus effects with intermediate allele frequencies. This was also found by Knott *et al* (1992a) for similar approximations. For parameter sets 1 and 2, both methods showed a large reduction of 35 to 40% for $r_{\mathbf{a},\hat{\mathbf{a}}}$ and 25 to 32% for $r_{\mathbf{Tg},\hat{\mathbf{Q}}\hat{\mathbf{g}}}$ when genotypes were unknown rather than known (tables II and III).

With the method of Kinghorn *et al* (1993), estimates of the allele frequency p were generally closer to the true values than with the other 2 procedures. However, major locus effects were overestimated and the correlations between true and predicted breeding values were close to zero which is in agreement with their simulation results. The method attempts to correct for the bias inherent in major locus estimates by regression on the independent variable $\mathbf{ZQ}^r$, an estimate from the data, which is associated with some error. The term $\mathbf{ZQ}^r$ is postmultiplied by the correction matrix $\mathbf{W}^r$. $\mathbf{ZQ}^r\mathbf{W}^r$ is then used the same way as a usual incidence matrix in the mixed model equations. Multiplication by $\mathbf{W}^r$ increases the variance of the independent variable to the variance expected for the unknown term $\mathbf{ZT}$. Because $\mathbf{W}^r$ is calculated over all animals with records, the new variance is correct only on the average. For an animal with known genotype, the elements in $\mathbf{Q}^r$ are identical to the values in $\mathbf{T}$ and should therefore not be altered by $\mathbf{W}^r$. Sires had more progeny than dams, therefore their estimated genotype probabilities were closer to the true values and should have been multiplied by a matrix

closer to an identity matrix in comparison to dams. In addition, breeding values estimated by [15] are still biased. These 2 problems are probably responsible for the overestimation of g and very poor prediction of polygenic breeding values. The performance of the method was, however, less affected by smaller allele effects (parameter set 3) than the other 2 procedures.

For all 3 procedures there was a problem of different solutions with different starting values when genotypes were unknown. For AML and the method of Hoeschele (1988) the cause could be the multimodality of the likelihood function. It seems to be necessary to compute approximated likelihoods which then can be used to select the solutions with the highest likelihood. This could of course also be done with the method of Kinghorn *et al* (1993) but this method has no direct relationship with maximum likelihood.

In this study variance components were assumed to be known but in practice have to be estimated. Using incorrect values could lead to biased estimates of major genotype effects and frequencies. For example, using an underestimated genetic variance might result in an overestimation of the major genotype effects. If a major allele is known to be segregating variance components free of major genotype effects would have to be estimated with model [1]. This could be very difficult because even when the true variance components were used, all 3 methods performed poorly when no animals were genotyped.

Clearly, none of the methods is satisfactory for a separate genetic evaluation for the major locus and the polygenes. In this study only large effects were considered. AML and, especially, the method of Hoeschele (1988) were unable to detect smaller effects than used with parameter set 3. For example, the effects estimated for the prolactin locus in a Holstein sire family (Cowan *et al*, 1990) were much smaller than considered here. The method proposed has some potential for improvement. Future research should focus on the development of algorithms to estimate genotype probabilities without any restriction on pedigree structures. The estimation of joint genotype probabilities for any 2 pairs of animals together with sparse matrix techniques to compute the elements of **M** could avoid the need for some of the approximations made in this study.

ACKNOWLEDGMENT

This research was conducted while AH was a visiting scientist at the University of Guelph. Financial support from the Schweizerischer Nationalfonds, Switzerland, is gratefully acknowledged.

REFERENCES

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

Eikelenboom G, Minkema D, van Eldik P, Sybesma W (1980) Performance of Dutch Landrace pigs with different genotypes for the halothane-induced malignant hyperthermia syndrome. *Livest Prod Sci* 7, 317-324

Elsen JM, Le Roy P (1989) Simplified versions of segregation analysis for detection of major genes in animal breeding data. *40th Annual Meeting of EAAP*, Dublin, 27-31 August, 1989

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. *Hum Hered* 21, 523-542

Fernando RL, Stricker C, Elston RC (1993) Scheme to compute the likelihood of a pedigree without loops and the posterior genotypic distribution for every member of the pedigree. *Theor Appl Genet*, in press

Henderson CR (1984) *Applications of Linear Models in Animal Breeding.* University of Guelph, Guelph, Canada

Hoeschele I (1988) Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. *Theor Appl Genet* 76, 81-92

Hoeschele I, Meinert TR (1990) Association of genetic defects with yield and type traits: The weaver locus effect on yield. *J Dairy Sci* 73, 2503-2515

Johnston J (1984) *Econometric Methods.* McGraw-Hill, New York

Kennedy BW, Quinton M, van Arendonk JAM (1992) Estimation of effects of single genes on quantitative traits. *J Anim Sci* 70, 2000-2012

Kinghorn BP, Kennedy BW, Smith C (1993) A method of screening for genes of major effect. *Genetics* 134, 351-360

Knott SA, Haley CS, Thompson R (1992a) Methods of segregation analysis for animal breeding data: a comparison of power. *Heredity* 68, 299-312

Knott SA, Haley CS, Thompson R (1992b) Methods of segregation analysis for animal breeding data: parameter estimates. *Heredity* 68, 313-320

Lange K, Elston RC (1975) Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Hum Hered* 25, 95-105

Le Roy P, Elsen JM, Knott S (1989) Comparison of four statistical methods for detection of a major gene in a progeny test design. *Genet Sel Evol* 21, 341-357

Le Roy P, Naveau J, Elsen JM, Sellier P (1990) Evidence for a new major gene influencing meat quality in pigs. *Genet Res Camb* 55, 33-40

MacLennan DH, Phillips MS (1992) Malignant hyperthermia. *Science* 256, 789-794

Morton NC, MacLean CJ (1974) Analysis of family resemblance. III. Complex segregation of quantitative traits. *Am J Hum Genet* 26, 489-503

Ott J (1979) Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *Am J Hum Genet* 31, 161-175

Van Arendonk JAM, Smith C, Kennedy BW (1989) Method to estimate genotype probabilities at individual loci in farm livestock. *Theor Appl Genet* 78, 735-740