Theor Appl Genet (1992) 83:878–886

THEORETICAL AND APPLIED GENETICS

A mixed model for analyses of data on multiple genetic markers

M. E. Goddard *

Center for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, Canada

Received January 11, 1991; Accepted June 21, 1991
Communicated by J. S. F. Barker

Summary. Data on a genetic marker linked to a gene affecting an important trait could help us to estimate breeding values for that trait more accurately. The accuracy is enhanced if many genetic markers are used and if important genes are bracketed by two markers. A mixed model for analysis of this type of data is presented. The model is appropriate for an arbitrary pedigree structure in an outbreeding species. It uses a "relationship" matrix among marked chromosome segments or QTL alleles. By using an analysis analogous to a reduced animal model, the number of effects to be estimated can be greatly reduced. A grouping strategy that can account for cross-breeding and linkage disequilibrium between markers and QTL alleles is included in the model. For analyses of a cross between inbred lines the model can be simplified. This simplification shows clearly the relationship of the mixed model analyses to multiple regression models used previously. The simplified model may also be useful for some experiments in outbreeding populations.

Key words: Genetic marker – Mixed models – BLUP

Introduction

If we could identify a genetic marker closely linked to a gene affecting an important character, it would allow us to select more accurately for that character. The possible advantages of this marker assisted selection (MAS) have been examined by Soller and Beckman (1983) and Smith and Simpson (1986). However, a single random marker is unlikely to be closely linked to a particular important gene. To make systematic use of MAS we will need to examine many markers so that any important gene will be closely linked to at least one marker.

Fortunately, recombinant DNA techniques are now providing a potentially unlimited number of markers. Consequently, MAS is already being used in plant improvement (Nienhuis and Helentjaris 1989) and experiments are commencing in animals.

Three broad methods of analysis for data on genetic markers and quantitative traits have been proposed.

1. Multiple regression models with one term for each marker.
2. Maximum likelihood.
3. Mixed model or best linear unbiased prediction (BLUP) (Fernando and Grossman 1989).

When multiple regression or least squares is used, the predicted breeding value of the best animals is over estimated. Smith and Simpson (1986) pointed this out in the context of MAS. Maximum likelihood methods that treat the effects to be estimated as fixed effects will also have this disadvantage. In addition, they tend to be difficult to compute for general data structures, i.e., other than nuclear human families and inbred line crosses.

For data that does not contain markers, BLUP has proven to be a very flexible method and one that does not overestimate the merit of the animals with highest estimated breeding value. BLUP can handle data with many nongenetic effects (e.g., herd), with arbitrary pedigree structure, and with nonrandom mating and selection. For these reasons it is also likely to be useful for analyzing data containing information on genetic markers, if the assumptions of BLUP are reasonably satisfied.

Fernando and Grossman (1989) demonstrated how this could be done for data on a single marker locus. However, in practice we are likely to generate data with

* On leave from: Livestock Improvement Unit, Department of Agriculture, PO Box 500, E. Melbourne 3002, Australia

many linked marker loci. Ideally, we would hope that important genes for quantitative traits were bracketed between two markers. The first aim of the current paper is to extend the model of Fernando and Grossman to deal with this situation.

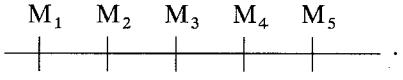
One important use of MAS is the introgression of genes from a resources population into a commercial line (Soller and Beckman 1983). This implies crossbreeding. Recently, Lande and Thompson (1990) have proposed that linkage disequilibrium between important genes and markers could be of significant value in MAS. The mixed model proposed in this paper accommodates crossbreeding and linkage disequilibrium.

In the case of a cross between inbred lines the model can be simplified. The simplified version of the model shows the relationship of the BLUP model to multiple regression approaches. The simplified model may be useful with non-inbred animals in some circumstances and this point is considered in the Discussion.

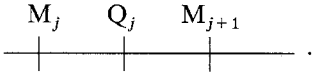
Basic model

The development of the model parallels that of Fernando and Grossman (1989) and uses similar notation.

Consider a chromosome with a series of marked loci



I will assume that there is at most one quantitative trait locus (QTL) between each pair of markers



At the locus Q_j each animal has two alleles, one inherited from its sire and one from its dam. Associated with each QTL allele is a marker haplotype consisting of the marker alleles at M_j and M_{j+1}. If the j^{th} chromosome segment that animal i inherited from its sire is of marker haplotype (kl), denote the value of the QTL allele on that segment by $v_{ij(kl)}$ or simply as $v_{ij(p)}$. Similarly, the allele from its dam is $v_{ij(m)}$. Then the breeding value of animal $i(a_i)$ summed over all chromosome segments is

$$a_i=\sum_j v_{ij(p)}+\sum_j v_{ij(m)}+u_i\,, \tag{1}$$

where u_i = breeding value at QTLs not included in marked segments.

The usual BLUP model for the phenotypic value of animal $i(y_i)$ is

$$y_i=\boldsymbol{x}_i'\boldsymbol{f}+a_i+e_i\,, \tag{2}$$

where

$\mathbf{f}$ = a vector of fixed effects,
$\boldsymbol{x}_i$ = an incidence vector,
e_i = environment deviation.

(Symbols for vectors are in bold and for matrices in uppercase bold.) On substituting Eq. 1 into Eq. 2 this becomes

$$y_i=\boldsymbol{x}_i'\boldsymbol{f}+\sum_j v_{ij(p)}+\sum_j v_{ij(m)}+u_i+e_i \tag{3}$$

or using matrix notation and assuming one record per animal

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{f}+\sum_j \boldsymbol{Z}_j\boldsymbol{v}_j+\boldsymbol{u}+\boldsymbol{e}\,, \tag{3a}$$

where the vector $\boldsymbol{v}_j$ contains two unknowns for each animal for each locus (one paternal and one maternal QTL effect)

the matrix $\boldsymbol{Z}_j$ has rows that contain two 1's and are otherwise zero; summation is over QTL loci (=chromosome segments bounded by markers)

To form the BLUP equations we require the covariance matrices of the **u**'s and of the **v**'s.

$$\mathrm{var}(\mathbf{u})=\mathbf{A}\sigma_u^2\,,$$

where

$\mathbf{A}$ = numerator relationship matrix,
σ_u^2 = variance of breeding value not associated with marked chromosome segments.

The variance of $\boldsymbol{v}$ is block diagonal, with each block corresponding to one QTL provided the base population is in linkage equilibrium. Similarly, the covariance of **u** and **v** is zero if the base population is in linkage equilibrium.

$$\mathrm{var}(\mathbf{e})=\mathbf{I}\boldsymbol{\sigma}^2$$

The BLUP equations are

$$\begin{pmatrix} X'X & X' & X'Z_1 & X'Z_2 & \cdots \\ X & I+A^{-1}\lambda & Z_1 & Z_2 & \cdots \\ Z_1'X & Z_1' & Z_1'Z_1+G_1^{-1}\boldsymbol{\sigma}^{-2} & Z_1'Z_2 & \cdots \\ Z_2'X & Z_2' & Z_2'Z_1 & Z_2'Z_2+G_2^{-1}\boldsymbol{\sigma}^{-2} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \begin{matrix} \hat{f} \\ \hat{u} \\ \hat{v}_1 \\ \hat{v}_2 \end{matrix} = \begin{matrix} X'y \\ y \\ Z_1'y \\ Z_2'y \end{matrix}\,, \tag{4}$$

where

$$\lambda = \frac{\sigma^2}{\sigma_u^2}$$

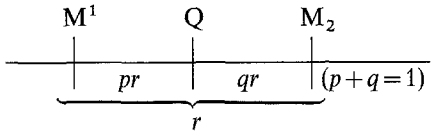
$$G_j = \mathrm{var}(v_j)$$

the number of **v** vectors, **Z** and **G** matrices is equal to the number of marked QTL with Eq. 4 logically extended. $\mathbf{Z}_j$ refers to the j^{th} marked QTL as in Eq. 3a.

A derivation of $\mathbf{G}_j^{-1}$ is given in the next section.

The var(v) matrix and $\mathbf{G}^{-1}$

Consider a single QTL bracketed by two marker loci with map distances as follows



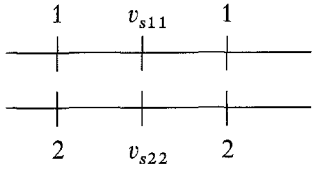
Assuming no interference between crossovers, the recombination rates are (Haldane 1919) between

M_1 and M_2 $a = 0.5(1 - e^{-2r})$

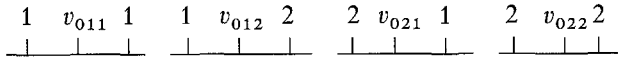
M_1 Q $b = 0.5(1 - 3^{-2pr})$

Q M_2 $c = 0.5(1 - e^{-2qr})$.

The relationships between the v's determine var(**v**), just as the relationships between animals determine **A**. The "parents" of the allele $v_{ij(p)}$ are the two alleles at that locus in the sire of i. Similarity, $v_{ij(m)}$ is derived from the two alleles in the dam. Without loss of generality the sires' genotypes can be writen as



Based on marker haplotypes, he will produce four types of gamete



The frequency and means of these four types of gamete are given in Table 1.

The approximate means are those obtained assuming no double recombinations between the markers. For the (1 1) and (2 2) haplotypes, the maximum error in the approximations can be shown to occur at $p = q = \frac{1}{2}$, where the true means are

$$(1\,1)\quad \left[1 - \frac{r^2}{4} + 0(r^2)\right] v_{s11} + \left[\frac{r^2}{4} + 0(r^2)\right] v_{s22}$$

$$(2\,2)\quad \left[\frac{r^2}{4} + 0(r^2)\right] v_{s11} + \left[1 - \frac{r^2}{4} + 0(r^2)\right] v_{s22},$$

Table 1. Frequencies and means of marker haplotypes

Haplotype		Frequency	Mean	Approximate mean
1	1	$\frac{1}{2}(1-a)$	$\frac{(1-b)(1-c)}{1-a} v_{s11} + \frac{bc}{1-a} v_{s22}$	v_{s11}
1	2	$\frac{1}{2}a$	$\frac{(1-b)c}{a} v_{s11} + \frac{b(1-c)}{a} v_{s22}$	$q v_{s11} + p v_{s22}$
2	1	$\frac{1}{2}a$	$\frac{b(1-c)}{a} v_{s11} + \frac{(1-b)c}{a} v_{s22}$	$p v_{s11} + q v_{s22}$
2	2	$\frac{1}{2}(1-a)$	$\frac{bc}{1-a} v_{s11} + \frac{(1-b)(1-c)}{1-a} v_{s22}$	v_{s22}

where 0(x) is a function such that $\frac{0(x)}{x}$ approaches zero as x approaches zero.

For the (1 2) and (2 1) haplotypes, the maximum error in the approximate means can be shown to occur at $p = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$, where the means are

$$(1\,2)\quad \left[p \pm \frac{r^2}{30} + 0(r^2)\right] v_{s11} + \left[q \pm \frac{r^2}{30} + 0(r^2)\right] v_{s22}$$

$$(2\,1)\quad \left[q \pm \frac{r^2}{30} + 0(r^2)\right] v_{s11} + \left[p \pm \frac{r^2}{30} + 0(r^2)\right] v_{s22}.$$

Thus, the value of the QTL in each gamete can be written in terms of the parental QTLs. For one representative of each haplotype and using the approximate means, these are

$$\begin{pmatrix} v_{011} \\ v_{012} \\ v_{021} \\ v_{022} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ q & p \\ p & q \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_{s11} \\ v_{s22} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \end{pmatrix} \tag{5}$$

The ε_{ij} are the deviation of each gamete from the mean of the haplotype. With the approximate mean $\varepsilon_{11} = \varepsilon_{22} = 0$ because v_{011} is identical to v_{s11} and v_{022} is identical to v_{s22}.

Following Fernando and Grossman (1989) and Quaas (1988) this method of writing QTL effects in terms of parental QTL alleles can be generalized for all QTL alleles in the data. That is,

$$\mathbf{v} = \mathbf{P}\mathbf{v} + \boldsymbol{\varepsilon}, \tag{6}$$

where **P** is a matrix similar to that in Eq. 5 and each row contains at most two nonzero terms which sum to 1. Thus,

$$(\mathbf{I} - \mathbf{P})\quad \mathbf{v} = \boldsymbol{\varepsilon}$$

$$\mathbf{v} = (\mathbf{I} - \mathbf{P})^{-1} \boldsymbol{\varepsilon}$$

$$\mathbf{G} = \mathrm{Var}(\mathbf{v}) = (\mathbf{I} - \mathbf{P})^{-1}\, \mathrm{var}(\boldsymbol{\varepsilon})\, (\mathbf{I} - \mathbf{P})'^{-1}$$

This allows us to find the inverse of G

$$\mathbf{G}^{-1} = [\mathrm{Var}(\mathbf{v})]^{-1} = (\mathbf{I} - \mathbf{P})' [\mathrm{var}(\boldsymbol{\varepsilon})]^{-1} (\mathbf{I} - \mathbf{P})$$

var(ε) is diagonal as shown in Appendix 1 and **P** is of simple structure, so that rules for $\mathbf{G}^{-1}$ can be derived in a similar way to that of Fernando and Grossman (1989). The rules are:

(1) Replace v_{011} with v_{s11} in all equations and delete rows and column of $\mathbf{G}^{-1}$ corresponding to v_{011}. Similarly, replace v_{022} with v_{s22}.
(2) For an offspring allele v_{012}

Add	to	Element of $\mathbf{G}^{-1}$
$\frac{1-p}{2p}$		(S11, S12)
$\frac{p}{2(1-p)}$		(S22, S22)
$\frac{1}{2p(1-p)}$		(012, 012)
$\frac{-1}{2p}$		(S11, 012) and (012, S11)
$\frac{-1}{2(1-p)}$		(S22, 012) and (012, S22)
$\frac{1}{2}$		(S11, S22) and (S22, S11)

(3) For an offspring allele v_{021} replace p with q and 012 and 021 in the above rules
(4) For an allele v_{s11} without known parents add 1 to element (S11, S11)

The main practical disadvantage of the approximation used for **P** is that, if double crossover does occur, $\hat{v}_{s11}$ and $\hat{v}_{011}$ are forced to be identical no matter how much evidence there is to the contrary. Therefore, it might be desirable to use a correlation sligthly less than 1 between v_{s11} and v_{011}. To do this, use

$$v_{011}=\left(1-\frac{r^2}{4}\right)v_{s11}+\frac{r^2}{4}v_{s22}+\varepsilon_{11}.$$

If this is adopted, row and columns for v_{011} are retained in $\mathbf{G}^{-1}$ and, in the rules given above, 012 is replaced by 011 and p is replaced by $r^2/4$.

Reduced animal model (RAM)

The Eqs. 4 are for a full animal model. If there are n QTL loci, there are $2n$ $\hat{v}$ effects and one $\hat{u}$ effect to be estimated for every animal. The number of effects to be estimated could be greatly reduced by use of a reduced animal model (Quaas and Pollock 1980). In a RAM, the breeding values of animals that are not parents are expressed in terms of their parents' breeding value. That procedure is now applied to QTL effects; that is, **v** effects are expressed in terms of their parental alleles just as they were in the derivation of $\mathbf{G}^{-1}$. The model for data on animals that are not parents becomes

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{f}+\boldsymbol{u}+\sum_j(\boldsymbol{P}_j^*\,\boldsymbol{v}_j+\boldsymbol{\varepsilon}_j)+\boldsymbol{e} \quad (7)$$
$$=\boldsymbol{X}\boldsymbol{f}+\boldsymbol{u}+\sum\boldsymbol{P}_j^*\,\boldsymbol{v}_j+\boldsymbol{e}^*,$$

where $e^*=\sum\varepsilon_j+e$

$\mathbf{P}_j^*$ = rows of $\mathbf{P}_j$ corresponding to animal that are not parents. The BLUP equations now contain $\hat{\boldsymbol{v}}$ terms only for animals that are parents.

Since the most efficient designs for estimating QTL effects involve a large number of offspring per sire, this should represent a substantial saving. A more detailed derivation of RAM for the original Fernando and Grossman (1989) model is given by Cantet and Smith (1991).

An example

As an illustration of the methods described above, consider the simple pedigree in Fig. 1. Genotypes at two linked marker loci are also given and these can be used to define alleles at a QTL located between the markers. It will be assumed that there are no double crossovers ($v_{011}=v_{s11}$) and for simplicity of presentation, fixed effects and genetic effects at other QTLs will be ignored. For this example, Eq. 3a becomes

$$\mathbf{y}=\mathbf{Z}\mathbf{v}+\mathbf{e}$$

i.e.,

$$\begin{pmatrix}y_s\\y_d\\y_o\end{pmatrix}=\begin{pmatrix}1&1&0&0&0\\0&0&1&1&0\\1&0&0&0&1\end{pmatrix}=\begin{pmatrix}v_{s11}\\v_{s22}\\v_{d33}\\v_{d44}\\v_{034}\end{pmatrix}+\begin{pmatrix}e_s\\e_d\\e_o\end{pmatrix}$$

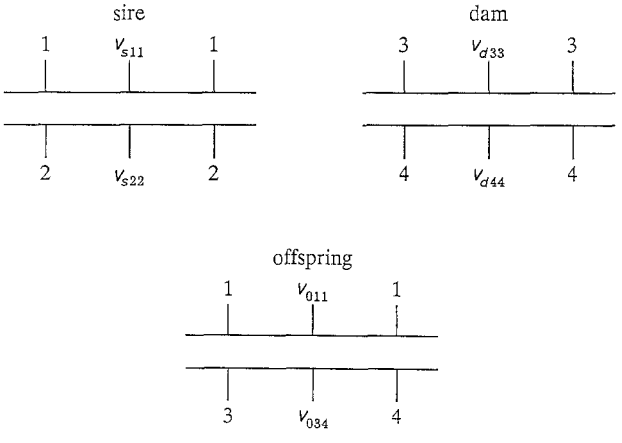


Fig. 1. Haplotypes of animals in a simple pedigree

The relationship matrix among the **v** is

$$G=\text{var}(\mathbf{v})$$

$$=\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 \\ & & 1 & 0 & q \\ \text{symmetric} & & & 1 & p \\ & & & & 1 \end{pmatrix}\sigma_v^2$$

Using the rules for inverting G gives

$$G^{-1}=\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 \\ & & 1+\frac{q}{2p} & \frac{1}{2} & -\frac{1}{2p} \\ \text{symmetric} & & & 1+\frac{p}{2q} & -\frac{1}{2q} \\ & & & & \frac{1}{2pq} \end{pmatrix}\sigma_v^2$$

For the reduced **v*** vector $\mathbf{G}=\text{var}(\mathbf{v}^*)=\mathbf{I}\,\sigma_v^2$, $\mathbf{G}^{-1}=\mathbf{I}\sigma_v^{-2}$ but now

$$\boldsymbol{R}=\text{var}(\boldsymbol{e}^*)=\begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2+2pq\sigma_v^2 \end{pmatrix}$$

and $\boldsymbol{R}^{-1}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \partial \end{pmatrix}\sigma^{-2}$

where $\partial=(1+2pq\lambda^{-1})^{-1}$

$$=\frac{\lambda}{\lambda+2pq}.$$

The mixed model equations for the RAM are $(\mathbf{Z}'\,\mathbf{R}^{-1}\,\mathbf{Z}+\mathbf{G}^{-1})\,\mathbf{v}=\mathbf{Z}'\,\mathbf{R}^{-1}\,\mathbf{y}$ i.e.,

$$\begin{bmatrix} 1+\partial+\lambda & 1 & q\partial & p\partial \\ 1 & 1+\lambda & & \\ q\partial & & 1+q^2\partial+\lambda & 1+pq\partial \\ p\partial & & 1+pq\partial & 1+p^2\partial+\lambda \end{bmatrix}\begin{pmatrix} \hat{v}_{s11} \\ \hat{v}_{s22} \\ \hat{v}_{d33} \\ \hat{v}_{d44} \end{pmatrix}=\begin{pmatrix} y_s+\partial y_o \\ y_s \\ y_d+q\partial y_o \\ y_d+p\partial y_o \end{pmatrix}$$

The mixed model Eqs. 4 become

$$\begin{pmatrix} 2+\lambda & 1 & & & 1 \\ 1 & 1+\lambda & & & \\ & & 1+\left(1+\frac{q}{2p}\right)\lambda & 1+\frac{1}{2}\lambda & -\frac{1}{2p}\lambda \\ & & 1+\frac{1}{2}\lambda & 1+\left(1+\frac{p}{2q}\right)\lambda & -\frac{1}{2q}\lambda \\ 1 & & -\frac{1}{2p} & -\frac{1}{2q}\lambda & 1+\frac{1}{2pq}\lambda \end{pmatrix}\begin{pmatrix} v_{s11} \\ v_{s22} \\ v_{d33} \\ v_{d44} \\ v_{034} \end{pmatrix}=\begin{pmatrix} y_s+y_o \\ y_s \\ y_d \\ y_d \\ y_o \end{pmatrix} \qquad (8)$$

where $\lambda=\frac{\sigma^2}{\sigma_v^2}$.

In the RAM, v_{034} is replaced $q\,v_{d33}+p\,v_{d44}+\varepsilon$ so that the model becomes

$$\begin{pmatrix} y_s \\ y_d \\ y_o \end{pmatrix}=\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & q & p \end{pmatrix}\begin{pmatrix} v_{s11} \\ v_{s22} \\ v_{d33} \\ v_{d44} \end{pmatrix}+\begin{pmatrix} e_s \\ e_d \\ e_o+\varepsilon \end{pmatrix}$$

i.e., $\mathbf{y}=\mathbf{Z}^*\,\mathbf{v}^*+\mathbf{e}^*$.

If in Eq. 8 for the full model, the $\hat{v}_{034}$ term is eliminated by absorbing the last row and column, the resulting equations are the same as those obtained above for the RAM. This is an example of the general rule that the full animal model and the RAM yield the same results.

Crossbreeding and linkage disequilibrium

So far we have assumed that the only relationships among the v effects are those due to the relationships

between animals included in the data. Two other causes of covariance are likely to occur.

(1) If there is linkage disequilibrium between markers and QTLs, all QTL alleles in the population bracketed by the same marker haplotype will be similar.
(2) QTL alleles that derive from the same breed will also be similar to the extent that breeds differ in mean value.

These effects can be incorporated into the model by including the equivalent of group effects. That is,

$$v_{ij(kl)} = b + L_j + h_{j(kl)} + d_{ij(kl)} ,$$

where

b = a breed effect common to all QTL alleles derived from that breed,
L_j = a breed effect common to all QTL alleles at the j^{th} locus derived from that breed,
$h_{j(kl)}$ = an effect common to all QTL alleles at the j^{th} locus that are contained in the kl marker haplotype.

The most useful grouping structure is probably that proposed by Thompson (1979), Quaas and Pollock (1981), Robinson (1986), and Westell et al. (1988). Applied to this case, the breed effect for a particular QTL allele is a combination of the breed effects of its parents.

E.g.:

$$b_{012} = q\, b_{s11} + p\, b_{s22} ,$$

where the subscripts denote animals as in previous sections.

This is equivalent to grouping the foundation alleles that do not have parents in the data (Westell et al. 1988). That is, for foundation alleles

$$v_{ij(kl)} = b + L_j + h_{j(kl)} + \varepsilon .$$

The **P** matrix used in Eq. 6 is augmented to reflect this so that

$$\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{L} \\ \boldsymbol{h} \\ \boldsymbol{v} \end{pmatrix} = \begin{pmatrix} \mathbf{0}' \\ \mathbf{0}' \\ \mathbf{0}' \\ \boldsymbol{P} \end{pmatrix} \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{L} \\ \boldsymbol{h} \\ \boldsymbol{v} \end{pmatrix} + \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{L} \\ \boldsymbol{h} \\ \boldsymbol{\varepsilon} \end{pmatrix}$$

i.e., $\mathbf{w} = \mathbf{P}^+ \mathbf{w} + \varepsilon^+$,

where

$$\mathbf{w} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{L} \\ \boldsymbol{n} \\ \boldsymbol{v} \end{pmatrix}, \mathbf{P}^+ = \begin{pmatrix} \mathbf{0}' \\ \mathbf{0}' \\ \mathbf{0}' \\ \boldsymbol{P} \end{pmatrix}, \varepsilon^+ = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{L} \\ \boldsymbol{h} \\ \boldsymbol{\varepsilon} \end{pmatrix}$$

$\mathbf{0}'$ = a row vector of zeros.

The inverse of var(**w**) is derived as before from

$$(\boldsymbol{I} - \boldsymbol{P}^+)' [\mathrm{var}(\varepsilon^+)]^{-1} (\boldsymbol{I} - \boldsymbol{P}^+) .$$

The rules for forming $[\mathrm{var}(\mathbf{w})]^{-1}$ are the same as previously described, except now for each foundation allele additions are made to the rows and columns representing groups. The group effects can be treated either as fixed or random by assigning the appropriate variances in $\mathrm{var}(\varepsilon^+)$. The BLUP equations are expanded to contain rows and columns for **b**, **I**, and **h** by adding zero columns to $\mathbf{Z}_i'$ (Westell et al. 1988).

In the derivation of the BLUP Eqs. 4 it was assumed that there was linkage equilibrium between the QTL. In this section, linkage disequilibrium between a QTL and its bracketing marker loci has been allowed for, but disequilibrium among the QTL themselves has not been dealt with. This should not be a major deficiency because, unless the QTL are tightly linked, they are not likely to display disequilibrium. An exception to this generalization is a population derived from crossing breeds or lines in which the QTL alleles from one breed are generally superior to those from the other. The **b** and **L** group affects help to overcome this problem, because then the d effects are deviations from the breed mean and so their correlation between loci will be reduced.

A simplification for inbred lines

The model describd above can be applied to any pedigree structure in an outbreeding population which can include crossbreeding. For a cross between inbred lines the model can be considerably simplified.

Consider a cross between inbred lines to produce an F_1 that is backcrossed to one of the parent lines. For data on the backcross generation, Eq. 3a becomes

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{f} + \sum_j \boldsymbol{Z}_j \boldsymbol{v}_j + \boldsymbol{u} + \boldsymbol{e} .$$

Only one $\mathbf{Z}_j \mathbf{v}_j$ term is needed per locus because the gametes from the other parent are all identical.

The equivalent RAM is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{f} + \sum_j \boldsymbol{P}_j \boldsymbol{v}_j + (\boldsymbol{u} + \sum_j \boldsymbol{\varepsilon}_j + \boldsymbol{e}) ,$$

where rows of $\mathbf{P}_j$ are as given in Eq. 5. $\mathbf{v}_j$ now contains only the two alleles per locus from the inbred parents. Only the difference between these alleles is estimable, so we can replace

v_{js11} by $M_j - v_j$
v_{js22} by $M_j + v_j$.

The M_j get absorbed into the overall mean which is part of $\mathbf{X}\mathbf{f}$.

Considering only one QTL, the model becomes

$$\mathbf{y}=\mathbf{X}\,\mathbf{f}+\mathbf{P}\,\mathbf{v}+\mathbf{e}^*, \tag{9}$$

where $\mathbf{e}^*=\mathbf{u}+\boldsymbol{\varepsilon}+\mathbf{e}$.

Writing one example of each haplotype and assuming no double crossovers, this is

$$\begin{pmatrix} y_{11}\\ y_{12}\\ y_{21}\\ y_{22}\end{pmatrix}=\boldsymbol{X}\boldsymbol{f}+\begin{pmatrix} -1\\ -(1-2p)\\ (1-2p)\\ +1\end{pmatrix} v+\boldsymbol{e}^*$$

Mixed model equations derived from Eq. 9 require that p be known. In the equivalent full model, p appears only in the $\mathbf{G}^{-1}$ matrix. It would be logical to estimate p using REML, then use the estimate in Eq. 9.

Alternatively Eq. 9 can be written as

$$\begin{pmatrix} y_{11}\\ y_{12}\\ y_{21}\\ y_{22}\end{pmatrix}=\boldsymbol{X}\boldsymbol{f}+\begin{pmatrix} -1 & 0\\ 0 & -1\\ 0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix} v\\ (1-2p)\,v\end{pmatrix}+\boldsymbol{e}^* \tag{10a}$$

$$=\boldsymbol{X}\boldsymbol{f}+\boldsymbol{Z}\,\boldsymbol{\theta}+\boldsymbol{e}^*. \tag{10b}$$

For each locus there is only one v and one p to be estimated, so we can treat $(1-2p)\,v$ as a single random variable. Assuming that v is normally distributed with mean 0 and variance σ_v^2 and that p is uniformly distributed from 0 to 1,

$$\mathrm{var}(\theta)=\mathrm{var}\begin{pmatrix} v\\ (1-2p)\,v\end{pmatrix}$$

$$=\begin{pmatrix} 1 & 0\\ 0 & 1/3\end{pmatrix}\sigma_v^2, \tag{10c}$$

as shown in Appendix 2.

Mixed model equations based on Eq. 10 can now be setup without prior estimation of p.

A second equivalent model is

$$\begin{pmatrix} y_{11}\\ y_{12}\\ y_{21}\\ y_{22}\end{pmatrix}=\boldsymbol{X}\boldsymbol{f}+\begin{pmatrix} -1 & -1\\ -1 & 1\\ 1 & -1\\ 1 & 1\end{pmatrix}\begin{pmatrix} (1-p)\,v\\ pv\end{pmatrix}+\boldsymbol{e}^* \tag{11a}$$

$$=\boldsymbol{X}\boldsymbol{f}+\boldsymbol{Z}_2\,\boldsymbol{\theta}_2+\boldsymbol{e}^*. \tag{11b}$$

As θ_2 is a linear transformation of θ

$$\mathrm{var}(\theta_s)=\mathrm{var}\begin{pmatrix} (1-p)\,v\\ pv\end{pmatrix}$$

$$=\begin{pmatrix} 1/3 & 1/6\\ 1/6 & 1/3\end{pmatrix}\sigma_v^2. \tag{11c}$$

For each marker locus, $\mathbf{Z}_2$ contains -1 when marker allele 1 is present and $+1$ when allele 2 is present.

Therefore Eq. 11 a, b, c is simply a model for regression on the markers with the regression coefficients (θ_2) treated as random effects with a certain covariance structure.

Now consider five linked markers

M_1	Q_1	M_2	Q_2	M_3	Q_3	M_4	Q_4	M_5	
1	$-v_1$	1	$-v_2$	1	$-v_3$	1	$-v_4$	1	1 parent 1 genotype
2	v_1	2	v_2	2	v_3	2	v_4	2	2 parent 2 genotype

The statistical model 11 can be expanded to

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{f}+\boldsymbol{Z}_3\,\boldsymbol{\theta}_3+\boldsymbol{e}^*, \tag{12}$$

where

$\theta_3'=(q_1v_1\ p_1v_1\ q_2v_2\ p_2v_2\ q_3v_3\ p_3v_3\ q_4v_4\ p_4v_4)$

$q_i=1-p_i$.

$\mathrm{var}(\theta_3)$ is block diagonal with each block as Eq. 11 c.

The model appears overparameterized, as it is a regression model with five variables (=markers) but eight unknowns. However, because the θ_s' are random variables, solutions are still possible.

Consider the row of $\mathbf{z}$ for the i^{th} animal denoted by $\mathbf{z}_i'$ with marker haplotype

$$\boldsymbol{z}_i'=\begin{matrix} 1 & \overbrace{1} & \overbrace{2} & \overbrace{2} & 1\\ -1 & -1\ -1 & 1\ 1 & 1\ 1 & -1\end{matrix}$$

$\boldsymbol{z}_i'$ contains two identical elements for each of the second, third, and fourth markers.

Consequently the model can be collapsed to

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{f}+\boldsymbol{Z}_4\,\boldsymbol{\theta}_4+\boldsymbol{e}, \tag{13}$$

where

$$\theta_4=\begin{pmatrix} q_1v_1\\ p_1v_1+q_2v_2\\ p_2v_2+q_3v_3\\ p_4v_4\end{pmatrix}$$

$$\mathrm{var}(\theta_4)=\begin{pmatrix} 1/3 & 1/6 & & & \\ 1/6 & 2/3 & 1/6 & & \\ & 1/6 & 2/3 & 1/6 & \\ & & 1/6 & 2/3 & 1/6\\ & & & 1/6 & 1/3\end{pmatrix}$$

This is now a multiple regression on the markers except that marker effects are treated as random. It requires that we know the order of markers along the chromosome so that $\mathrm{var}(\theta_4)$ can be specified correctly. If this order were unknown, the approximation

$\mathrm{var}(\theta_4)=\boldsymbol{I}\,\sigma_c^2$ could be used.

If instead, we treat the θ_4 as fixed effects, we have the standard multiple regression model.

Discussion

The basic model proposed here is very similar to that of Fernando and Grossman (1989). Even the rules for inverting G look similar to their rules, although p (position of QTL within marker bracket) in this paper has a quite different meaning to d (recombination rate between QTL and marker) in theirs. However, a major advance in accuracy of estimating breeding values is achieved by having markers bracketing a QTL. This is seen in the **G** matrix by the near identity of QTL effects in parent and offspring if they share the same marker haplotype. Consequently, information on the value of chromosome segments does not erode so quickly from one generation to the next. This is important since large amounts of data are needed to estimate these effects with any precision.

The grouping strategy suggested is by no means the only one available, but it illustrates how biological phenomena can be represented in the statistical model. Beckman and Soller (1988) discussed the identification of QTL's in a cross between non-inbred lines. They assumed that the lines were fixed for alternate alleles at the QTL, possibly as a result of past selection. Under these conditions, the cross of outbred lines was almost as efficient as an inbred lines cross in estimating QTL effects (Beckman and Soller 1988). In the model proposed here, this situation is represented by the L_j term, which groups the alleles at the j^{th} locus that are derived from one breed. If breeds are fixed at the QTL, deviations from L_j ($d_{ij\,(kl)}$) would be set to zero. The L_j effect could be treated as fixed, but it is probably better to treat them as random if a number of L_j effects must be estimated. If L_j is treated as random, the b effects should be included in the model.

MAS is useful in introgressing a few desirable genes from a resource strain into a commercial line (Soller and Beckman 1983). It helps to retain the few desired genes and to recover the commercial genotype at all other loci. The b group effect is important in achieving the latter. It groups all chromosome segments derived from the same breed. If only b was included in the model of breeding value, the effect would be to rank animals on the proportion of the genome derived from each breed. That is, animals with the largest number of chromosome segments from the superior breed would be given the highest estimated breeding value. By including L_j and $d_{ij\,(kl)}$ in the model, this simple estimate is modified as detailed data on loci (L_j) and individual alleles ($d_{ij\,(kl)}$) accumulate.

Animals that carry the same marker haplotype are likely to be related and therefore to carry the same QTL allele. If the relationship is included in the data, this will be reflected in the **G** matrix. However, the relationship may have occurred before the base generation of the data set. The $h_{(kl)}$ grouping term recognizes this. The linkage disequilibrium that it implies may be the result of previous crossbreding (Lande and Thompson 1990) or simply of finite population size.

The mixed model equations require estimates of the terms in the **G** matrix i.e., p and σ_v^2 for each locus. A Baysian approach, which starts with prior estimates and modifies them in the light of new information, seems appropriate. As a priori, one might assume that the total genetic variance (σ_g^2) is divided equally among the QTLs. Then, if there were 150 marker brackets $2\sigma_v^2$ would be $\sigma_g^2/150$. Thus σ_v^2 is very small and $\hat{v}$ estimates will be regressed back considerably unless the amount of data on which they are based is large. This is not a deficiency of the BLUP analysis but a recognition that the effects at the "average" QTL must be small. However, it is probably important to allow σ_v^2 to vary between loci, so that genuinely major loci can be recognized and treated accordingly (M. Goddard, W. Zhang, C. Smith, in preparation).

As well as estimates of σ_v^2, the basic model requires estimates of p for each QTL. The simplified model for inbred line crosses avoids this problems by incorporating p into the random effect to be estimated. This would cause the $\Theta = (1-2p)\,v$ effects not to be normally distributed, but this is unlikely to cause a serious problem. It is only strictly valid if there is only one Θ effect to be estimated for each **p**, as is the case in a cross of inbred lines. This is also the case for an experiment in an outbreeding species consisting of a large number of half-sib offspring from one sire. In this case, the sire is like the F_1 parent of a backcross and the variation in breeding value coming from the dams becomes part of the error variance. However, in a more complex pedigree structure there will be one effect (v_j) for each allele at the locus, but the same **p** will apply to all of them. Consequently, treating each $(1-2p)\,v_i$ as an independent random variable would lose some information. However, the loss may not be great. For instance, consider an experiment in which a large number of half-sib offspring are produced from each of three sires. Model 13 could be applied to each sire's offspring. The error term (e^*) now includes the effect of QTL alleles inherited from the dams.

The simplified models shows clearly the relationship of the BLUP analysis to multiple regression. In this case the only difference is treating the QTL effects associated with markers as random variables. This avoids the tendency of multiple regression to exaggerate the merit of the best animals. Also because the $\mathbf{G}^{-1}$ matrix is not proportional to the $\mathbf{Z}'\mathbf{Z}$ matrix, BLUP does not regress all regression coefficients equally. Consequently, the BLUP analysis ranks animals in a different order to the multiple regression analysis. In more complex data structure, the advantage of the BLUP analysis would be greater as it accounts for relationships, selection, and varying amounts of information on different QTL alleles.

The derivation of BLUP does not require that the random effects in the model be normally distributed. However, some desirable properties of BLUP do require normality. The distribution of v in Eq. 3a is not expected to be normally distributed. Goddard et al. (1991), using computer simulation, found that marked departure from normality in the distribution of QTL effect did not decrease the accuracy of estimated breeding value and that BLUP performed better than multiple regression.

A full maximum likelihood analysis would use these departure from normality and should therefore be more accurate, provided it also treated the v's as random effects. However, the gain in accuracy may be small and the flexibility and generality of the BLUP approach should make it useful. The BLUP approach presented here can cope with any pedigree structure including crossbreeding in an outbreeding species. This prevents severe computational problems for a maximum likelihood analysis (Fernando 1990).

Appendix 1: Proof that var(ε) is diagonal

Assume that the QTL alleles in a parent s have value v_{s11} and v_{s22}. The value of the QTL allele inherited from this parent by offspring o is

$$v_{0kl}=\begin{cases} v_{s11} \text{ with probability } \varrho \\ v_{s22} \text{ with probability } 1-\varrho \end{cases}. \quad \text{(A1)}$$

For instance, if $v_{0kl}=v_{012}$, then using the approximate means in Table 1, $\varrho=q$.

In Eq. 5 this is expressed as a linear model

$$v_{0kl}=\varrho\, v_{s11}+(1-\varrho)\, v_{s22}+\varepsilon_{0kl}. \quad \text{(A2)}$$

Combining (A1) and (A2) the distribution of ε_{0kl} is

$$\varepsilon_{0kl}=\begin{cases} (1-\varrho)(v_{s11}-v_{s22}) & \text{with probability } \varrho \\ \varrho(v_{s22}-v_{s11}) & \text{with probability } 1-\varrho \end{cases}. \quad \text{(A3)}$$

Each gamete that is formed receives randomly and independently one of the parents QTL alleles, and so each ε is an independent realization from the distribution given by Eq. A3. Since each realization of ε is independent, they must be uncorrelated and so var(ε) is diagonal.

Appendix 2: Derivation of var(θ)

Assume v is normally distributed with mean 0 and variance σ_v^2, p is uniformly distributed from 0 to 1 and hence has mean 1/2 and variance 1/12, v and p are independent.

Then $\operatorname{var}[(1-2p)\,v]=E(1-2p)^2\,v^2-[E(1-2p)\,v]^2$
$=E(1-2p)^2\,E(v^2)-[E(1-2p)\,E(v)]^2$
$=E(1-4p+4p^2)\,E(v^2)-0$
$=(1-\frac{4}{2}+\frac{4}{3})\,\sigma_v^2$
$=\frac{1}{3}\,\sigma_v^2$.

Similarly,

$\operatorname{cov}[(1-2p)\,v, v]=E(1-2p)\,v^2-E(1-2p)\,v\,E(v)$
$=E(1-2p)\,E(v^2)-0$
$=0$.

References

Beckman JS, Soller M (1988) Detection of linkage between marker loci and loci affecting quantitative traits in crosses between segregating populations. Theor Appl Genet 76:228–236

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

Fernando RL (1990) Statistical problems in marker-assisted selection for QTL. Proc 4th World Congr Genet Appl Livestock Prod 13:433–436

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

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

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Nienhuis J, Helentjaris T (1989) Simultaneous selection for multiple polygenic traits through RFLP analyses. In: Development and application of molecular markers to problems in plant genetics. Cold Spring Habor Laboratory Press, Cold Spring Harbor/NY, pp 107–112

Quaas RL (1988) Additive genetic model with groups and relationships. J Dairy Sci 71:1338–1345

Quaas RL, Pollock EJ (1980) Mixed model methodoly for farm and ranch beef cattle testing programs. J Anim Sci 51:1277

Quaas RL, Pollock EJ (1981) Modified equations for sire models with groups. J Dairy Sci 64:1868

Robinson GK (1986) Group effects and computing strategies for models of estimated breeding values. J Dairy Sci 69:3106–3111

Smith C, Simpson SP (1986) The use of genetic polymorphism in livestock improvement. J Anim Breed Genet 103:205–217

Soller M, Beckman JS (1983) Genetic polymorphism in varietal identification and genetic improvement. Theor Appl Genet 67:25–33

Thompson R (1979) Sire evaluations. Biometrics 35: 339

Westell RA, Quaas RL, Van Vleck LD (1988) Genetic groups in an animal model. J Dairy Sci 71:1310–1318