ORIGINAL ARTICLE

Prediction of response to marker-assisted and genomic selection using selection index theory

J. C. M. Dekkers

Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, IA, USA

Keywords
inbreeding; genomic selection; marker assisted selection; selection index; selection response.

Correspondence
J. C. M. Dekkers, Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, 50011-3150, IA, USA. Tel: +515-294-7509; Fax: +1 515-294-9150;
E-mail: jdekkers@iastate.edu

Received: 7 March 2007;
accepted: 2 October 2007

Summary

Selection index methods can be used for deterministic assessment of the potential benefit of including marker information in genetic improvement programmes using marker-assisted selection (MAS). By specifying estimates of breeding values derived from marker information (M-EBV) as a correlated trait with heritability equal to 1, it was demonstrated that marker information can be incorporated in standard software for selection index predictions of response and rates of inbreeding, which requires specifying phenotypic traits and their genetic parameters. Path coefficient methods were used to derive genetic and phenotypic correlations between M-EBV and the phenotypic data. Methods were extended to multi-trait selection and to the case when M-EBV are based on high-density marker genotype data, as in genomic selection. Methods were applied to several example scenarios, which confirmed previous results that MAS substantially increases response to selection but also demonstrated that MAS can result in substantial reductions in the rates of inbreeding. Although further validation by stochastic simulation is required, the developed methodology provides an easy means of deterministically evaluating the potential benefits of MAS and to optimize selection strategies with availability of marker data.

Introduction

The recent availability of high-density marker maps and low costs of genotyping large numbers of markers using high-throughput genotyping methodology has renewed interests in incorporating marker information in programmes for genetic improvement of livestock through the use of marker-assisted selection (MAS). Of particular interest is the use of genomic selection, as proposed by Meuwissen *et al.* (2001), which uses associations of large numbers of markers across the genome with phenotypes, capitalizing on linkage disequilibrium (LD) between markers and closely linked quantitative trait loci (QTL), without prior screening of markers based on significance of their associations with the phenotype. The resulting predictions of the random effects of marker haplotypes (Meuwissen *et al.* 2001), or of alleles at each marker (Solberg *et al.* 2006), are then used to predict breeding values for individuals based on their genotype for all markers. By simultaneously selecting on large numbers of markers, this is in contrast to most strategies for MAS that have been used to date, which are based on a limited number of markers or genes (see review by Dekkers 2004).

Before incorporating markers in breeding programmes, careful assessment is required of the potential benefits of MAS and of the design of breeding programmes to optimally capitalize on the benefits of MAS. Neimann-Sorensen & Robertson (1961) and Smith (1967) were the first to propose that selection index theory can be used to incorporate information on individual loci into selection strategies. These selection index methods were later extended by Lande & Thompson (1990). Most recent work on evaluating the impact of information on individual genes or markers has, however, been based on stochastic rather than deterministic simulations (e.g. Verrier 2001) because deterministic prediction of response by selection index theory requires multi-variate normality, which is violated when genotypes of only a limited number of markers or genes are used in MAS (Lande & Thompson 1990). In addition, although methods to incorporate changes in genetic variance through selection-induced gametic phase disequilibrium (Bulmer effect Falconer & Mackay 1996) have been developed (Wray & Hill 1989; Villanueva *et al.* 1993), selection index predictions ignore changes in genetic variances that result from changes in allele frequencies, which will be substantial when high selection emphasis is placed on a limited number of loci. Assumptions of multi-variate normality of breeding values derived using genetic markers and small changes in gene frequencies will, however, be more valid if MAS is based on information from a large number of markers across the genome (Lande & Thompson 1990), as would be the case with genomic selection. Thus, the advent of genomic selection offers renewed opportunities for the use of selection index theory to deterministically evaluate and optimize the use of markers in selection programmes. Deterministic models have substantial advantages over stochastic simulations because they require much less computing time and are more amenable to optimization.

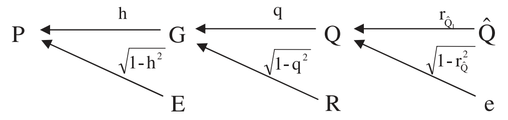
The main purpose of this study was, therefore, to formulate selection index methods for deterministic prediction of the potential benefit of MAS on response to selection and, in particular, to do this in a manner that facilitates use of standard selection index software that has been developed (e.g. Rutten *et al.* 2002). Resulting methodology provides an effective means for initial evaluation of MAS for implementation in industry programmes. Methods will be illustrated with examples that demonstrate the potential benefit of genomic selection in a limited number of scenarios.

Methods

Single trait MAS index formulation

With the availability of LD markers (Dekkers 2004), the total additive genetic value of an additive quantitative trait (G) can be partitioned into genetic effects that are correlated with markers through LD (Q) and residual genetic effects (R) that are independent of the markers. Note that, in addition to QTL that are not in LD with the markers, *R* also includes effects resulting from incomplete LD of QTL that are linked to the markers. This partitioning results in the following model for the phenotypes,

$$P = G + E = Q + R + E$$



P, phenotype;
G, total genetic value;
E, environmental component of *P*;
Q, component of *G* that is associated with markers;
R, component of *G* that is independent of markers;
$\hat{Q}$, estimated breeding value (EBV) for *Q* based on marker data;
e, prediction error of $\hat{Q}$;
h^2, heritability of *P*;
q^2, proportion of genetic variance associated with markers;
$r_{\hat{Q}}$, accuracy of $\hat{Q}$ as a predictor of *Q*;
r_{MG}, accuracy of $\hat{Q}$ as a predictor of *G*.

Figure 1 Path coefficient diagram illustrating the relationships among components contributing to phenotype with marker-assisted selection for a single trait.

where *E* represents random environmental effects. A path diagram of this model is in Figure 1. Note that when markers used for marker-assisted genetic evaluation are randomly located across the genome, as would be the case for genomic selection, effects included in *Q* and *R* represent a random partitioning of QTL effects into effects that are associated with markers through LD (=*Q*) and effects that are independent of marker genotypes (=*R*).

Let h^2 denote the total heritability for the trait, and q^2 the proportion of genetic variance contributed by *Q*. Proportion q^2 depends on the genetic variance contributed by QTL that are in LD with markers and the extent of LD between markers and QTL. For an individual QTL linked to a single marker, q^2 is equal to the product of LD between the marker and the QTL, as measured by r^2 (Hill & Robertson 1968), and the proportion of total genetic variance that is contributed by the QTL. When the QTL is in LD with multiple markers, q^2 will depend on the r^2 of the QTL with any of its surrounding markers and on the structure of LD around the QTL or on the r^2 of the QTL with multi-marker haplotypes (e.g. Hayes *et al.* 2006).

When individuals are genotyped for markers that are in LD with QTL across the population (LD markers), marker effects can be estimated across families from an analysis of phenotype and marker genotype data obtained from the population. Here, it will be assumed that estimates are obtained from fitting markers or haplotypes as random rather than fixed effects, i.e. they represent estimates of breeding values with the properties of Best Linear Unbiased Prediction (BLUP: Henderson 1984), similar to BLUP estimated breeding values (EBV) derived from phenotypes. Such a model with markers as random effects was described by Meuwissen *et al.* (2001) for genomic selection, with estimates of marker effects derived from phenotypic data and high-density single nucleotide polymorphism (SNP) genotypes in one generation, which were then used to obtain marker-based EBV (M-EBV) of individuals based on their marker genotypes for several subsequent generations. When based on multiple regions of the genome, or on all regions of the genome, as with genomic selection, the M-EBV of an individual can be computed as the sum of estimates across alleles or haplotypes for each genomic region j as:

$$\hat{Q} = \sum_j (\hat{g}_j^{pat} + \hat{g}_j^{mat})$$

where $\hat{g}_j^{pat}$and $\hat{g}_j^{mat}$ are the BLUP estimates of the effects of the paternal and maternal marker alleles or marker haplotypes for interval j.

M-EBV, $\hat{Q}$, are estimates of genetic effects Q. Using properties of BLUP EBV (Henderson 1984), the relationship between M-EBV and Q can be modeled as: $Q = \hat{Q} + e$, where e represents the (unknown) prediction error for an individual's M-EBV (Figure 1). The model for the phenotype can then be expanded as:

$$P = \hat{Q} + e + R + E.$$

Note that the use of BLUP to estimate M-EBV results in a zero correlation between $\hat{Q}$ and its prediction error e (Henderson 1984), as reflected in Figure 1.

Let $r_{\hat{Q}}$ denote the accuracy of $\hat{Q}$ as a predictor Q, i.e. the correlation between Q and $\hat{Q}$. Then, path coefficients associating Q, $\hat{Q}$ and the prediction error e can be derived and are presented in Figure 1. The correlation of $\hat{Q}$ with G is equal to $r_{MG} = qr_{\hat{Q}}$. This correlation represents the accuracy of the M-EBV as a predictor of the total genetic value G, and represents the accuracies of M-EBV for genomic selection that were obtained by Meuwissen *et al.* (2001). The proportion of genetic variance that is explained by the M-EBV then is equal to r^2_{MG}, which is equivalent to parameter p defined by Lande & Thompson (1990).

Unless all QTL that affect the trait have been identified, selection on M-EBV must be combined with selection on any available phenotypic information, to ensure simultaneous improvement of both Q and R. To accommodate this and following Neimann-Sorensen & Robertson (1961), Smith (1967), and Lande & Thompson (1990), marker and phenotypic information can be combined in an index of the following form:

$$I = \mathbf{b}'\mathbf{X} = \left[\mathbf{b}'_Q, \mathbf{b}'_P\right] \begin{bmatrix} \mathbf{X}_Q \\ \mathbf{X}_P \end{bmatrix}$$

where $\mathbf{X}_Q$ is a vector with M-EBV on the individual itself and/or its relatives, $\mathbf{X}_P$ is a vector with phenotypic records on the individual itself and/or its relatives, and $\mathbf{b}_Q$ and $\mathbf{b}_P$ are vectors of index weights. Lande & Thompson (1990) showed that index weights could be derived by standard selection index methodology (Hazel 1943) for predicting the overall genetic value G, using

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_Q \\ \mathbf{b}_P \end{bmatrix} = \mathbf{P}^{-1}\mathbf{G}$$
$$= \begin{bmatrix} \mathrm{Var}(\mathbf{X}_Q) & \mathrm{Cov}(\mathbf{X}_Q, \mathbf{X}'_P) \\ \mathrm{Cov}(\mathbf{X}_P, \mathbf{X}'_Q) & \mathrm{Var}(\mathbf{X}_P) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{Cov}(\mathbf{X}_Q, \mathbf{G}) \\ \mathrm{Cov}(\mathbf{X}_P, \mathbf{G}) \end{bmatrix}$$

and the corresponding accuracy of selection as: $r_{G,I} = \sqrt{\frac{\mathbf{b}'\mathbf{G}}{\sigma^2_G}}$, which can then be used to predict response to selection with intensity i based on i $\mathrm{r}_{G,I}\sigma_G$ (Falconer & Mackay 1996). Following Lande & Thompson (1990) and using the property of BLUP that the covariance of BLUP EBV with true breeding values is equal to the variance of BLUP EBV (Henderson 1984), elements ij of all matrices and vectors involving $\mathbf{X}_Q$ are equal to $a_{ij}r^2_{MG}\sigma^2_G$, where a_{ij} is the additive genetic relationship between individuals or groups represented by column i and row j. Matrices and vectors that do not involve $\mathbf{X}_Q$ are obtained by using standard quantitative genetics theory based on phenotypic data (Falconer & Mackay 1996). These methods can also be extended to include data on multiple traits and multiple trait breeding goals, as demonstrated by Lande & Thompson (1990).

Implementation of MAS indexes

Selection index methods for MAS proposed by Lande & Thompson (1990) are not immediately in a form that is suitable for use in standard selection index procedures, such as the program SelAction (Rutten *et al.* 2002), which require specification of phenotypic traits with their heritabilities, standard deviations and phenotypic and genetic correlations as input parameters. It is, however, possible to accommodate marker information in these programs by specifying M-EBV as a correlated trait with heritability equal to 1 and using appropriate correlations between the original trait and the trait M-EBV, as will be demonstrated in the following. Inclusion of QTL information in SelAction as a trait with unit heritability was also used by Schrooten et al. (2005) but only for a single-trait situation and is further justified in the discussion. Methods similar to those described here were recently also used to evaluate MAS for commercial crossbred performance (Dekkers, 2007).

Using the path coefficient diagram in Figure 1, the following correlations that are required for inclusion of M-EBV as a trait in selection index calculations can be derived [see Lynch & Walsh (1998) for a recent description of path coefficient theory]. The genetic correlation between the original trait and the trait M-EBV is: $r_{G\hat{Q}} = qr_{\hat{Q}} = r_{MG}$. The corresponding phenotypic correlation is: $r_{P\hat{Q}} = hqr_{\hat{Q}} = hr_{MG}$. Together with a heritability of the trait M-EBV of 1 and a phenotypic (and therefore genetic) standard deviation of M-EBV of $r_{\mathrm{MG}}\sigma_G$, these parameters result in variances and covariances that are identical to the elements in matrix **P** and vector **G** of the Lande & Thompson (1990) derivation. For example, the covariance of the 'phenotype' for M-EBV of individual or group i with phenotype for the original trait of the individual or the group j (=element of matrix **P**) is equal to the genetic covariance between these variables (=element of vector **G**), because a heritability of 1 is used for M-EBV and can be derived using standard quantitative genetics theory for covariances between observations on correlated traits (Falconer & Mackay 1996), as: $\mathrm{Cov}(\hat{Q}_i, P_j) = \mathrm{Cov}(\hat{Q}_i, G_j) = a_{ij}\mathrm{Cov}(\hat{Q}, G) = a_{ij}r_{G\hat{Q}}\sigma_G r_{MG}\sigma_G = a_{ij}r^2_{MG}\sigma^2_G$. Thus, this formulation of marker information allows parameters to be entered in a trait-based form into standard selection index procedures and software.

Extension to multiple traits

The purpose of this section is to extend the developed methodology to multiple traits by deriving appropriate correlations between the traits involved, i.e. phenotype-based traits and marker-EBV-based traits. The general theory that will be developed also applies to cases in which a selected group of markers is used for selection, e.g., as determined based on prior QTL studies. Use of genomic selection will, however, result in several simplifying assumptions because of the random associations of markers with QTL across the genome, and this will be presented as a special case throughout the following.

Let $\rho_{G_{12}}$, $\rho_{R_{12}}$, and $\rho_{Q_{12}}$ be the genetic correlations between traits 1 and 2 for the genetic components G, R, and Q. The partitioning of genetic effects into Q and R results in Q_1 to be uncorrelated to R_1 and Q_2 uncorrelated to R_2. Genetic correlations between Q_1 and Q_2 and between R_1 and R_2 are expected to be equal to the genetic correlation between G_1 and G_2 $(E(\rho_{R_{12}}) = \mathrm{E}(\rho_{Q_{12}}) = \rho_{G_{12}})$ if the same markers are used for MA-genetic evaluation for both traits and if markers have not been pre-selected based on associations with the phenotype. This is expected to hold for genomic selection because genetic effects associated with markers will then be comprised of a random proportion of genetic effects that contribute to each trait. For other cases, these correlations need to be estimated. The correlation between environmental components contributing to traits 1 and 2 is denoted by $\rho_{E_{12}}$.

Using the path coefficient diagram in Figure 2, the following phenotypic and genetic correlations between the phenotypic and the marker-based traits that are necessary for derivation of selection indices can then be determined:

$$r_{G_i\hat{Q}_j} = q_i r_{\hat{Q}_i} r_{\hat{Q}_i\hat{Q}_j} + q_i\sqrt{1 - r^2_{\hat{Q}_i}}\, r_{\hat{Q}_j e_i} = q_i r_{\hat{Q}_i} r_{\hat{Q}_i} r_{\hat{Q}_j}\rho_{Q_{12}}$$
$$+ q_i\sqrt{1 - r^2_{\hat{Q}_i}}\, r_{\hat{Q}_i}\sqrt{1 - r^2_{\hat{Q}_i}}\,\rho_{Q_{12}} = q_i r_{\hat{Q}_j}\rho_{Q_{12}};$$

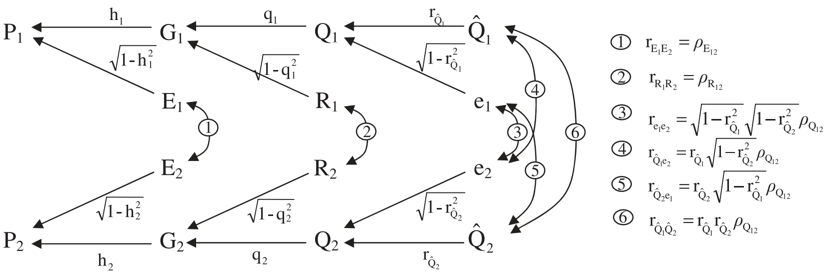
$$r_{P_i\hat{Q}_j} = h_i r_{G_i\hat{Q}_j} = h_i q_i r_{\hat{Q}_i}\rho_{Q_{12}}.$$

With random allocation of markers, the proportion of genetic variance that is associated with markers, q_i^2, is expected to be equal for both traits, in which case correlations simplify to:

$$r_{G_i\hat{Q}_j} = r_{MG_j}\rho_{Q_{12}};$$

$$r_{P_i\hat{Q}_j} = h_i r_{MG_j}\rho_{Q_{12}}.$$

Correlations can be further simplified by replacing $\rho_{Q_{12}}$ by $\rho_{G_{12}}$. These parameters and their simplifications under genomic selection are summarized in Table 1. Note that the accuracy of M-EBV as a predictor of Q_i, and therefore also the accuracy of M-EBV as a predictor of G_i $(r_{MG_i} = q_i r_{\hat{Q}_i})$, can differ


P_1
h_1
G_1
q_1
Q_1
$r_{\hat{Q}_1}$
$\hat{Q}_1$
$\sqrt{1-h_1^2}$
$\sqrt{1-q_1^2}$
$\sqrt{1-r_{\hat{Q}_1}^2}$
E_1
R_1
e_1
1
2
3
4
5
6
E_2
R_2
e_2
$\sqrt{1-h_2^2}$
$\sqrt{1-q_2^2}$
$\sqrt{1-r_{\hat{Q}_2}^2}$
P_2
G_2
Q_2
$\hat{Q}_2$
h_2
q_2
$r_{\hat{Q}_2}$
1 $r_{E_1E_2}=\rho_{E_{12}}$
2 $r_{R_1R_2}=\rho_{R_{12}}$
3 $r_{e_1e_2}=\sqrt{1-r_{\hat{Q}_1}^2}\sqrt{1-r_{\hat{Q}_2}^2}\rho_{Q_{12}}$
4 $r_{\hat{Q}_1e_2}=r_{\hat{Q}_1}\sqrt{1-r_{\hat{Q}_2}^2}\rho_{Q_{12}}$
5 $r_{\hat{Q}_2e_1}=r_{\hat{Q}_2}\sqrt{1-r_{\hat{Q}_1}^2}\rho_{Q_{12}}$
6 $r_{\hat{Q}_1\hat{Q}_2}=r_{\hat{Q}_1}r_{\hat{Q}_2}\rho_{Q_{12}}$

P_i, phenotype for trait i;
G_i, total genetic component of P_i;
E_i, environmental component of P_i;
Q_i, component of G_i that is associated with markers;
R_i, component of G_i that is independent of markers;
$\hat{Q}_i$, EBV for Q_i based on marker data;
e_i, prediction error of $\hat{Q}_i$;
h_i^2, heritability of P_i;
q_i^2, proportion of genetic variance associated with markers for trait i;
$r_{\hat{Q}_i}$, accuracy of $\hat{Q}_i$ as a predictor of Q_i;
r_{MG_i}, accuracy of $\hat{Q}_i$ as a predictor of G_i;
$\rho_{G_{12}}$, genetic correlation between traits 1 and 2;
$\rho_{P_{12}}$, phenotypic correlation between traits 1 and 2;
$\rho_{Q_{12}}$, correlation between Q_1 and Q_2;
$\rho_{R_{12}}$, correlation between residual genetic effects for traits 1 (R_1) and 2 (R_2).

Figure 2 Path coefficient diagram illustrating the relationships among traits and genetic components with multi-trait marker-assisted selection.

Table 1 Genetic parameters[1] for four traits considered for derivation of selection criteria: phenotype for trait 1 (P_1) and trait 2 (P_2), and marker-based estimated breeding values (EBV) for trait 1 ($\hat{Q}_1$) and trait 2 ($\hat{Q}_2$)

	P_1	P_2	$\hat{Q}_1$	$\hat{Q}_2$
P_1	h_1^2	$\rho_{P_{12}}$	$h_1q_1r_{\hat{Q}_1} =^2 h_1r_{MG_1}$	$h_1q_1r_{\hat{Q}_2}\rho_{Q_{12}} = h_1r_{MG_2}\rho_{G_{12}}$
P_2	$\rho_{G_{12}}$	h_2^2	$h_2q_2r_{\hat{Q}_1}\rho_{Q_{12}} = h_2r_{MG_1}\rho_{G_{12}}$	$h_2q_2r_{\hat{Q}_2}$
$\hat{Q}_1$	$q_1r_{\hat{Q}_1} = r_{MG_1}$	$q_2r_{\hat{Q}_1}\rho_{Q_{12}} = r_{MG_1}\rho_{G_{12}}$	1	$r_{\hat{Q}_1}r_{\hat{Q}_2}\rho_{Q_{12}} = r_{\hat{Q}_1}r_{\hat{Q}_2}\rho_{G_{12}}$
$\hat{Q}_2$	$q_1r_{\hat{Q}_2}\rho_{Q_{12}} = r_{MG_2}\rho_{G_{12}}$	$q_2r_{\hat{Q}_2} = r_{MG_2}$	$r_{\hat{Q}_1}r_{\hat{Q}_2}\rho_{Q_{12}} = r_{\hat{Q}_1}r_{\hat{Q}_2}\rho_{G_{12}}$	1

[1] h_i^2, heritability of the phenotype for trait i;
q_i^2, proportion of genetic variance associated with the markers for trait i;
$r_{\hat{Q}_i}$, accuracy of $\hat{Q}_i$ as a predictor of marker-associated genetic effects, Q_i;
r_{MG_i}, accuracy of $\hat{Q}_i$ as a predictor of the total genetic value, G_i;
$\rho_{G_{12}}$, genetic correlation between traits 1 and 2;
$\rho_{P_{12}}$, phenotypic correlation between traits 1 and 2;
$\rho_{Q_{12}}$, correlation between Q_1 and Q_2;
$\rho_{R_{12}}$, correlation between residual genetic effects for traits 1 (R_1) and 2 (R_2).

[2] Results after the equality signs apply to genomic selection and assume $q_1 = q_2$ and $\rho_{G_{12}} = \rho_{Q_{12}} = \rho_{R_{12}}$, and use $q_ir_{\hat{Q}_i} = r_{\mathrm{MG}_i}$.

between traits because not only does it depend on the amount of phenotypic data but also on the accuracy, i.e. heritability, of the phenotypic data that is available to estimate marker effects.

The final correlation that is needed for multiple-trait selection on M-EBV is the correlation between the M-EBV for the two traits. Based on the assumption that phenotypic data that contribute to $\hat{Q}_1$ and $\hat{Q}_2$ are independent, which will underestimate the correlation if traits 1 and 2 are measured on the same animals but will approximately be true if sufficient data are used to estimate marker effects, these correlations can be derived to be equal to:

$$r_{\hat{Q}_1\hat{Q}_2} = \frac{\mathrm{Cov}(\hat{Q}_1, \hat{Q}_2)}{\sqrt{\mathrm{Var}(\hat{Q}_1)\mathrm{Var}(\hat{Q}_2)}} = \frac{r_{\hat{Q}_1}^2 r_{\hat{Q}_2}^2 \rho_{Q_{12}}}{r_{\hat{Q}_1} r_{\hat{Q}_2}} = r_{\hat{Q}_1} r_{\hat{Q}_2} \rho_{Q_{12}}$$

This correlation cannot be further simplified to depend only on r_{MG} and $\rho_{Q_{12}}$.

For the multiple-trait case, it should also be noted that, although the use of BLUP to estimate M-EBV results in a zero correlation between the M-EBV for a trait, $\hat{Q}_i$, and its prediction error, e_i, when M-EBV are obtained from single-trait procedures, which is what is assumed here, prediction errors of an individual's M-EBV for trait 1 (2) will be correlated to prediction errors of its M-EBV for trait 2 (1).

In addition, prediction errors for M-EBV for trait 1 (2) will also be correlated with the M-EBV for trait 2 (1). Using the previously derived correlations, these correlations can be found to equal:

$$r_{\hat{Q}_1 e_2} = \frac{\mathrm{Cov}(\hat{Q}_1, Q_2 - \hat{Q}_2)}{\sqrt{\mathrm{Var}(\hat{Q}_1)\mathrm{Var}(e_2)}} = \frac{\mathrm{Cov}(\hat{Q}_1, Q_2) - \mathrm{Cov}(\hat{Q}_1, \hat{Q}_2)}{\sqrt{\mathrm{Var}(\hat{Q}_1)\mathrm{Var}(e_2)}}$$

$$= \frac{r^2_{\hat{Q}_1}\rho_{Q_{12}} - r^2_{\hat{Q}_1} r^2_{\hat{Q}_2}\rho_{Q_{12}}}{r_{\hat{Q}_1}\sqrt{1 - r^2_{\hat{Q}_2}}} = r_{\hat{Q}_1}\sqrt{1 - r^2_{\hat{Q}_1}}\rho_{Q_{12}};$$

$$r_{\hat{Q}_1 e_1} = r_{\hat{Q}_2}\sqrt{1 - r^2_{\hat{Q}_1}}\rho_{Q_{12}};$$

$$r_{e_1 e_2} = \sqrt{1 - r^2_{\hat{Q}_1}}\sqrt{1 - r^2_{\hat{Q}_2}}\rho_{Q_{12}}.$$

Note that, using path diagram theory (Lynch & Walsh 1998) and the path diagram in Table 2, these correlations result in the correct correlation between Q_1 and Q_2:

$$r_{Q_1 Q_2} = r_{\hat{Q}_1} r_{\hat{Q}_1 \hat{Q}_2} r_{\hat{Q}_2} + r_{\hat{Q}_1} r_{\hat{Q}_1 e_2}\sqrt{1 - r^2_{\hat{Q}_2}} + \sqrt{1 - r^2_{\hat{Q}_1}} r_{\hat{Q}_2 e_1} r_{\hat{Q}_2} + \sqrt{1 - r^2_{\hat{Q}_1}} r_{e_1 e_2}\sqrt{1 - r^2_{\hat{Q}_2}},$$

which, when substituting the previous equations for correlations among EBV and prediction errors, simplifies to $\rho_{Q_{12}}$.

Incorporating the Bulmer effect and predicting rates of inbreeding

Reparameterizing the model by specifying M-EBV as correlated traits with heritability equal to 1, in principle allows methods that have been developed for predicting response to selection for polygenic traits to be applied to MAS. This includes pseudo-BLUP selection index methods for deterministic modelling of selection on Animal Model BLUP EBV (Wray & Hill 1989), incorporation of the effects of selection on variance-covariance structures through the Bulmer effect (Wray & Hill 1989; Villanueva *et al.* 1993) and effects of co-selection of relatives on selection intensities (Meuwissen 1991). In addition, methods developed for prediction of rates of inbreeding based on long-term contribution theory (Woolliams & Bijma 2000) can be used. Such methods have been implemented in the selection index software package SelAction (Rutten *et al.* 2002) and this software will be used in the following to demonstrate use of the model to predict potential benefits of MAS. When applying these methods and this software it must, however, be realized that use of a trait with heritability equal to 1 may push the validity of the developed methods and that all predictions are based on multi-variate normality and the infinitesimal model and, therefore, do not account for changes in gene frequencies. In addition, using M-EBV as a genetic trait in these methods assumes that the same estimates of marker or haplotype effects are used over generations, such that the composition of the M-EBV remains constant.

Examples of MAS index predictions

To illustrate the use of the developed methodology to predict the potential benefit of MAS, markers used were assumed to be randomly allocated across the genome, reflecting genomic selection, thus $\rho_{R_{12}} = \rho_{Q_{12}} = \rho_{G_{12}}$. This same assumption also causes the expected proportion of genetic variance that is associated with markers to be equal for all traits, thus $q_1 = q_2$, which leads to $q_i r_{\hat{Q}_i} = q_j r_{\hat{Q}_j} = r_{MG}$, which is the accuracy of the M-EBV as a predictor of the total genetic value. Under these assumptions and using r_{MG} as an input parameter, phenotypic and genetic correlations between phenotypes and marker-based EBV depend only on r_{MG}, and not on its partition into q and $r_{\hat{Q}}$.This makes the results more general and applicable to different combinations of q and $r_{\hat{Q}}$ for a given level of accuracy of M-EBV (r_{MG}). The correlation between M-EBV ($r_{\hat{Q}_1 \hat{Q}_2} = r_{\hat{Q}_1} r_{\hat{Q}_2}\rho_{\hat{Q}_{12}}$) does depend on $r_{\hat{Q}_i}$.

All calculations were performed using the program SelAction (Rutten *et al.* 2002) using pseudo-BLUP selection index procedures. Asymptotic responses to selection after reaching equilibrium values for variances and covariances based on the Bulmer effect are reported. Predictions of rates of inbreeding were as implemented in the program SelAction based on long-term contribution theory.

Results

Single trait MAS

Figures 1 and 2 show the impact of marker information on asymptotic response to selection for single-trait selection in a simplified pig breeding programme as a function of the accuracy of M-EBV (r_{MG} ranging from 0 to 1). Each generation, 20 males were selected. Each male was mated to three selected females, which each produced eight offspring (four males, four females). Heritability of the phenotypic trait was 0.4 (Figure 1) or 0.1 (Figure 2) and selection was on BLUP EBV based on phenotypic and/or marker data. The base scenario was a trait with phenotype recorded on all individuals prior to selection.

Results show that using markers alone will require accuracies of M-EBV of at least 0.75 when heritability of the trait is 0.4 (Figure 1) and of at least 0.55 when trait heritability is 0.1 (Figure 2). These are for cases when phenotype is recorded on all individuals prior to selection. Combined selection, using both phenotypic and marker data, resulted in substantial extra responses over phenotype-based selection, in particular for the trait with low heritability and when phenotypes were recorded on females only. Figures 1 and 2 also show that genotyping males only reduced extra responses from including marker data by 30–40%.

Although M-EBV was modeled as a trait with heritability equal to 1 and does, therefore, not benefit from including information on relatives, not including M-EBV data of relatives in the selection index did reduce overall response, as demonstrated in Figures 3 and 4. The reason is that M-EBV of relatives contribute to the evaluation of residual genetic effects, as was also demonstrated by Lande & Thompson (1990).

Figures 5 and 6 show the impact of including marker information on rates of inbreeding. Results show that selection on marker information alone can dramatically reduce rates of inbreeding, from 2% to over 3% per generation with phenotypic selection, to less than 1% per generation with selection on M-EBV alone. The reason for this is that markers provide information on the Mendelian sampling terms received by the individual and reduces the impact of pedigree information, which increases probabilities of co-selection of relatives. Thus, with use of marker information, emphasis is moved from between-family selection to within-family selection. When selecting on a combination of phenotypic and marker data, the impact of including marker information on reducing rates of inbreeding depends on the emphasis that is placed on the M-EBV and, therefore, on its accuracy, r_{MG}. When only males are genotyped, rates of inbreeding were substantially higher compared with genotyping all individuals (Figures 5 and 6).

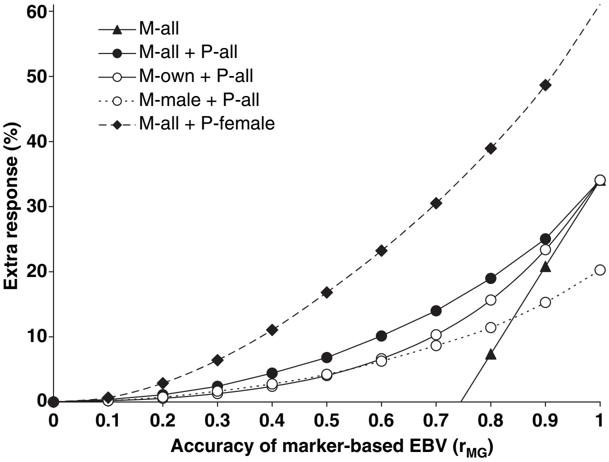


Figure 3 Effect of the accuracy of marker-based estimated breeding values (M-EBV) (r_{MG}) on extra response to selection (% over selection based on phenotypic data alone) for a trait with heritability equal to 0.4. Cases represented are selection on M-EBV alone (M-all) and combined selection on M-EBV and phenotype, with genotypes available on all individuals (M-all), the selection candidate alone (M-own), or males alone (M-males), and phenotypes available on all individuals (P-all) or only females (P-female).

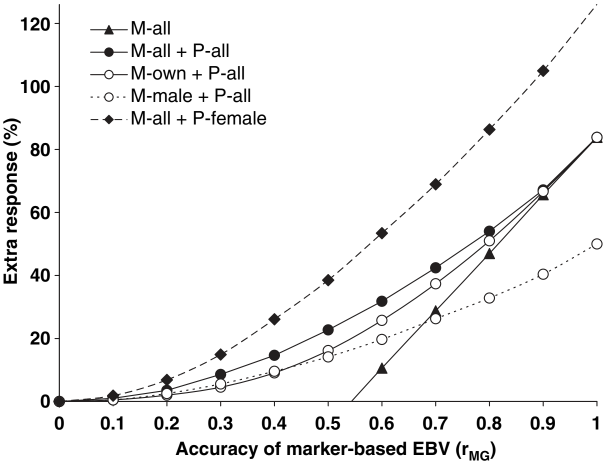


Figure 4 Effect of the accuracy of marker-based estimated breeding values EBV (M-EBV) (r_{MG}) on extra response to selection (% over selection based on phenotypic data alone) for a trait with heritability equal to 0.1. Cases represented are selection on M-EBV alone (M-all) and combined selection on M-EBV and phenotype, with genotypes available on all individuals (M-all), the selection candidate alone (M-own), or males alone (M-males), and phenotypes available on all individuals (P-all) or only females (P-female).

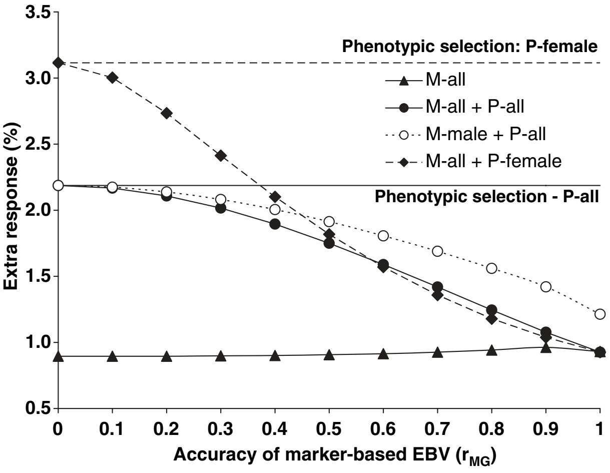


Figure 5 Effect of the accuracy of marker-based estimated breeding values EBV (M-EBV) (r_{MG}) on rates of inbreeding per generation for a trait with heritability 0.4. Cases represented are selection on phenotype alone, on M-EBV alone (M-all), and combined selection on M-EBV and phenotype, with genotypes available on all individuals (M-all) or on males alone (M-males), and phenotypes available on all individuals (P-all) or only females (P-female).

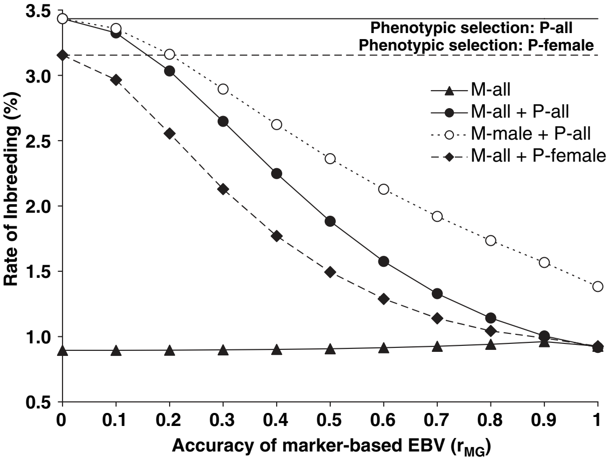


Figure 6 Effect of the accuracy of marker-based estimated breeding values EBV (M-EBV) (r_{MG}) on rates of inbreeding per generation for a trait with heritability 0.1. Cases represented are selection on phenotype alone, on M-EBV alone (M-all), and combined selection on M-EBV and phenotype, with genotypes available on all individuals (M-all) or on males alone (M-males), and phenotypes available on all individuals (P-all) or only females (P-female).

Table 2 Genetic parameters for selection on a breeding goal of two traits (P_1 and P_2) with and without marker information and resulting responses to selection in individual traits and the breeding goal (ΔH) and rates of inbreeding (ΔF). Marker-based estimated breeding values (EBV) ($\hat{Q}_1$ and $\hat{Q}_2$) have accuracies of 0.8, based on markers explaining 62.4% of the genetic variance

Correlations[1]	P_1	P_2	$\hat{Q}_1$	$\hat{Q}_2$	ΔH	ΔF(%)
P_1	–	−0.5	0.438	−0.131		
P_2	−0.3	–	−0.076	0.253		
$\hat{Q}_1$	0.8	−0.24	–	−0.243		
$\hat{Q}_2$	−0.24	0.8	−0.243	–		
Heritability	0.3	0.1	1	1		
Phenotypic SD	1	1	0.8	0.8		
Economic value	1	1	0	0		
Response to selection						
Phenotype only	0.408	0.041	0.394	0.052	0.448	2.36
Markers only	0.418	0.068	0.655	0.167	0.486	0.94
Combined	0.469	0.074	0.582	0.148	0.543	1.29

[1] Phenotypic correlations above the diagonal; genetic correlations below the diagonal SD, standard deviation.

Multiple trait MAS

Table 2 shows results for an example of multiple trait selection for a breeding goal with two negatively correlated traits. The structure of the population was the same as used in the single-trait example. All individuals were assumed to be phenotyped for the trait prior to selection. Trait 1 had moderate heritability (0.3) and was negatively correlated to trait 2, which had low heritability (0.1). Thus, this example could represent combined selection for growth and health or reproduction. Phenotypic and genetic correlations of phenotypic and M-EBV traits were based on equations in Table 1, assuming genomic selection with an accuracy of $r_{MG} = 0.8$ and that markers explain 62.4% of the genetic variance for each trait ($q^2 = 0.624$). The latter percentage is required only to compute the correlation between M-EBV for the two traits (Table 1).

Selection on markers alone resulted in 8.5% greater response in the breeding goal than selection on phenotype only (Table 2). Most of the extra gain resulted from a 66% greater response in the less heritable trait, which was difficult to improve by phenotypic selection because of its negative correlation with the more heritable trait. Selecting on the combination of phenotypic and marker data resulted in 21.2% greater gain in the breeding goal than phenotypic selection and in 80.5% greater response in the low heritable trait. Rates of inbreeding were reduced by nearly 50% with combined selection and even more with marker-only selection.

Discussion

The main purpose of this work was to present a formulation for data obtained from markers in a breeding programme that allows for the evaluation of the incorporation of marker information in selection strategies using pseudo-BLUP selection index theory and inbreeding prediction methodology. This was achieved by reparameterizing the selection index theory that was developed for MAS by Lande & Thompson (1990) by modelling M-EBV as a separate trait with heritability equal to 1. Methods and theories associated with BLUP prediction of breeding values and path coefficient theory were then used to derive the associated variances and covariances needed for trait-based selection indexes. This formulation provides a convenient basis for including marker information as a correlated trait in selection index calculations. The resulting methodology allows marker information to be included in standard selection index software such as SelAction (Rutten *et al.* 2002), as was also demonstrated by Schrooten et al. (2005) for single-trait selection. While methods presented can be used to model MAS on any set of markers of genes, several assumptions that are expected to be approximately valid with genomic selection were capitalized on to derive fairly straightforward formulations of correlations between phenotypic data and M-EBV for implementation in standard selection index programmes.

Formulating M-EBV as a trait with heritability one also allowed marker information from relatives to be incorporated in a natural way, which is needed to investigate the impact of not genotyping all individuals. With heritability equal to 1, correlations between M-EBV on relatives will be equal to the additive genetic relationship between relatives. Additive genetic relationships quantify the correlation between additive effects of relatives for single and multiple loci (Falconer & Mackay 1996) and can, therefore, also be used to model correlations of M-EBV between relatives, and to predict M-EBV of an individual based on the M-EBV of relatives.

Potential benefits from MAS

For illustration purposes, methods were applied to several examples, using the program SelAction (Rutten *et al.* 2002). Results demonstrated the increased genetic gains that can be achieved with availability of M-EBV but that, unless M-EBV have high accuracy, they should be used in conjunction with available phenotypic data. Similar to what has been demonstrated in numerous other simulation studies (see e.g. Dekkers & Hospital 2002), extra gains from MAS were greatest for cases where phenotypic data provides limited accuracy of selection, including sex-limited traits and traits with low heritability. The example of multi-trait selection demonstrated that MAS will be particularly beneficial for increasing response for economic traits for which responses are low in current phenotype-based selection programmes, because of limited accuracy of EBV or undesirable correlations with other economic traits with higher heritability. Similar results were observed by Verrier (2001) using stochastic simulation with a single QTL. Results also showed that MAS can result in substantial reductions in rates of inbreeding because of the increased emphasis on own rather than family information.

Only one-stage selection programmes were investigated here. Several studies have, however, shown that marker information is particularly beneficial in multi-stage selection programmes, where marker data is used in early stages when limited phenotypic data is available. This includes pre-selection of young bulls in dairy cattle for entry into progeny testing programmes, as was investigated by Kashi *et al.* (1990). Schaeffer (2006) proposed that, with genomic selection, early selection on M-EBV could remove the need for progeny testing in dairy cattle, thereby reducing generation intervals as well as costs. Methods described here for incorporating M-EBV could also be used to investigate such multiple-stage selection strategies.

Model assumptions

Selection index methods are based on several assumptions that are required for selection index predictions of responses to be valid. The most important one is the assumption of multivariate normality of M-EBV. It should be noted that the derivation of selection index weights does not require this assumption and can be applied even to MAS with one QTL or gene (e.g. Dekkers & Settar 2003). However, the use of selection index methods to predict response to selection and inbreeding does rely on the fundamental assumption of multivariate normality.

The multi-variate normal assumption will be approximately valid if M-EBV are based on a substantial number of markers or QTL regions (Lande & Thompson 1990), in which case the Central Limit theorem dictates an approximate normal distribution of M-EBV, thereby allowing them to be modelled as a polygenic trait. Although the validity of this assumption depends on the number of markers included in the M-EBV and on the distribution of the marker effects, it will be approximately valid for genomic selection. When based on multiple regions of the genome, or on all regions of the genome, as with genomic selection, the M-EBV of a progeny can be computed as the sum of estimates of effects on phenotype of alleles or haplotypes for each genomic region j as: $\hat{Q}_{progeny} = \sum_j(\hat{g}_j^{pat} + \hat{g}_j^{mat})$, as demonstrated previously. When the same estimates of allele or haplotype effects are used for several generations, the M-EBV of a progeny can also be written as the average of the M-EBV of its parents plus the sum of deviations for alleles or haplotypes that are transmitted to the progeny:

$$\hat{Q}_{progeny} = (1/2)\hat{Q}_{sire} + (1/2)\hat{Q}_{dam} + \left(\sum_j \hat{g}_{ij}^{pat} - (1/2)\hat{Q}_{sire}\right) + \left(\sum_j \hat{g}_{ij}^{mat} - (1/2)\hat{Q}_{dam}\right).$$

Note that this is equivalent to the Mendelian genetic model that is assumed for polygenic breeding values, with the latter two terms representing the Mendelian sampling terms. When based on multiple QTL regions and markers, these Mendelian sampling terms will approximately follow a normal distribution, which is what is assumed for polygenic traits. Further, because an individual's M-EBV is fixed conditional on marker genotypes and previously derived estimates of marker effects, it has no residual term. Thus, it can be observed without error based on the individual's marker genotypes and two individuals with the same marker genotypes will have the same M-EBV, hence the assumption of heritability equal to 1. Note that this does assume that (if needed) parental origin of alleles or haplotypes can be determined without error and that estimates of marker or haplotype effects remain consistent across several generations. Thus, although M-EBV represent estimates, they can be viewed and modeled as a genetic trait that is inherited in a polygenic manner and that can be observed on individuals without error (i.e. no environmental effect).

The model used for selection index predictions also assumed that the accuracy of M-EBV remains constant over generations, apart from the impact of the Bulmer effect on variances and covariances. If LD between markers is not complete, gene frequencies change substantially, or if dominance and epistatic effects play a role, marker-effects will need to be re-estimated on a regular basis to maintain accuracy. Using updated estimates of marker effect estimates, however, violates the assumption of M-EBV being a consistent trait across generations. The model also assumed that marker effects were estimated on phenotypic data that were independent of the phenotypic data that were used for phenotype-based EBV. This will be approximately true when using LD markers because marker effects will be estimated from a sample across families, thereby limiting the impact of individual families on marker estimates.

It is clear that, ultimately, the deterministic selection index predictions developed here must be validated by stochastic simulation. This was, however, beyond the scope of the present study because of the complexity of the simulation and genetic evaluation models that would be required but is the subject of ongoing research. Nevertheless, the developed models are based on the established theory that was validated under the infinitesimal model and should also apply with the use of M-EBV under the assumption of normality. The developed models, therefore, allow an initial assessment of the benefit of marker information and can provide the basis for further development of deterministic models for MAS that allow rapid assessment of alternate strategies of selection.

Correlations between phenotypic data and M-EBV that are required for incorporation of marker information were shown to only depend on the accuracy of the marker-based EBV, r_{MG}, which was used as an input parameter in the examples used here to illustrate methodology. Accuracy r_{MG} depends on the proportion of genetic variance explained by markers (q^2) and the accuracy of estimates of marker effects that are in LD with QTL, $r_{\hat{Q}}$. Both of these parameters are to some extent under the control of the breeder. With genomic selection, parameter q^2 depends on marker density and on the extent and pattern of LD that exists in the population. Parameter $r_{\hat{Q}}$ depends on the amount and accuracy of data available to estimate marker effects and on the efficacy of the statistical methods used for estimation or prediction. Deterministic methods to predict r_{MG} for a given marker density, LD structure, and amount of phenotypic information have not yet been developed, but they can be derived by stochastic simulation, as in Meuwissen *et al.* (2001). For the purposes of the work presented herein, r_{MG} was used as an input parameter and the effect of different levels of r_{MG} on responses was evaluated.

Acknowledgements

The author is grateful for discussions with Rohan Fernando on aspects of the work presented here and for financial support from the Iowa Agriculture and Home Economics Experiment Station, Ames, IA, USA (Project No. 3600) and Hatch Act and State of Iowa funds.

References

Dekkers J.C.M. (2004) Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J. Anim. Sci.*, **82**, E313–E328.

Dekkers J.C.M. (2007) Marker-assisted selection for commercial crossbred performance. *J. Anim. Sci.*, **85**, 2104–2114.

Dekkers J.C.M., Hospital, F. (2002) Utilization of molecular genetics in genetic improvement of plants and animals. *Nat. Rev.: Genet.*, **3**, 22–32.

Dekkers J.C.M., Settar P. (2003) Long-term selection with known quantitative trait loci. In: J. Janick (ed.), Plant Breeding Reviews, Vol. 24, Part 1, Long Term Selection: Maize. Wiley & Sons, Inc., New York, USA pp. 311–336.

Falconer D.S., Mackay T.F.C. (1996) Introduction to Quantitative Genetics. Longman, Harlow, UK.

Hayes B.J., Chamberlain, A.J., Goddard, M.E. (2006) Use of markers in linkage disequilibrium with QTL in breeding programs. Electronic communication no. 30-06. In: Proc. 8th World Congress on Genetics Applied to Livestock Production. Belo Horizonte, MG, Brazil: http://www.wcgalp8.org.br.

Hazel L.N. (1943) The genetic basis for constructing selection indices. *Genetics*, **38**, 476–490.

Henderson C.R. (1984) Applications of linear models in animal breeding. Univ. Guelph, Guelph, Ontario, Canada.

Hill W.G., Robertson A. (1968) Linkage disequilibrium in finite populations. *Theoret. Appl. Genet.*, **38**, 226–231.

Kashi Y., Hallerman E., Soller M. (1990) Marker-assisted selection of candidate bulls for progeny testing programmes. *Anim. Prod.*, **51**, 63–74.

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

Lynch M., Walsh B. (1998) Genetics and Analysis of Quantitative Traits. Sinauer Accoc. Inc., Sunderland, MA.

Meuwissen T.H.E. (1991) Reduction of selection differentials in finite populations with a nested full-half sib family structure. *Biometrics*, **47**, 195–203.

Meuwissen T.H.E., Hayes B., Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Neimann-Sorensen A., Roberson A. (1961) The association between blood groups and several production characters in three Danish cattle breeds. *Acta Agric. Scand.*, **11**, 163–196.

Rutten M.J.M., Bijma P., Woolliams J.A., van Arendonk J.A.M. (2002) SelAction: Software to predict selection response and rate of inbreeding in livestock breeding programs. *J. Heredity*, **93**, 456–458.

Schaefer L.R. (2006) Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.*, **123**, 218–223.

Schrooten C., Bovenhuis, H., van Arendonk, J.A.M. Bijma, P. (2005) Genetic progress in multistage dairy cattle breeding schemes using genetic markers. *J. Dairy Sci.*, **88**, 1569–1581.

Smith C. (1967) Improvement in metric traits through specific genetic loci. *Anim. Prod.*, **9**, 349–358.

Solberg T.R., Sonesson A., Woolliams J., Meuwissen T.H.E. (2006) Genomic selection using different marker types and density. Electronic communication no. 22-13. In: Proc. 8th World Congress on Genetics Applied to Livestock Production. Belo Horizonte, MG, Brazil: http://www.wcgalp8.org.br.

Verrier E. (2001) Marker assisted selection for the improvement of two antagonistic traits under mixed inheritance. *Genet. Sel. Evol.*, **33**, 17–38.

Villanueva B., Wray N.R., Thompson R. (1993) Prediction of asymptotic rates of response from selection on multiple traits using univariate and multivariate best linear unbiased predictors. *Anim. Prod.*, **57**, 1–13.

Woolliams J.A., Bijma P. (2000) Predicting rates of inbreeding in populations undergoing selection. *Genetics*, **154**, 1851–1864.

Wray N.R., Hill W.G. (1989) Asymptotic rates of response from index selection. *Anim. Prod.* **49**, 217–227.