Genet. Sel. Evol. 36 (2004) 395–414

DOI: 10.1051/gse:2004008

Original article

A study on the minimum number of loci required for genetic evaluation using a finite locus model

Liviu R. Totir[a*], Rohan L. Fernando[a,b], Jack C.M. Dekkers[a,b], Soledad A. Fernández[c]

[a] Department of Animal Science, Iowa State University, Ames, IA 50011, USA
[b] Lawrence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA
[c] Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

(Received 22 August 2003; accepted 22 March 2004)

Abstract – For a finite locus model, Markov chain Monte Carlo (MCMC) methods can be used to estimate the conditional mean of genotypic values given phenotypes, which is also known as the best predictor (BP). When computationally feasible, this type of genetic prediction provides an elegant solution to the problem of genetic evaluation under non-additive inheritance, especially for crossbred data. Successful application of MCMC methods for genetic evaluation using finite locus models depends, among other factors, on the number of loci assumed in the model. The effect of the assumed number of loci on evaluations obtained by BP was investigated using data simulated with about 100 loci. For several small pedigrees, genetic evaluations obtained by best linear prediction (BLP) were compared to genetic evaluations obtained by BP. For BLP evaluation, used here as the standard of comparison, only the first and second moments of the joint distribution of the genotypic and phenotypic values must be known. These moments were calculated from the gene frequencies and genotypic effects used in the simulation model. BP evaluation requires the complete distribution to be known. For each model used for BP evaluation, the gene frequencies and genotypic effects, which completely specify the required distribution, were derived such that the genotypic mean, the additive variance, and the dominance variance were the same as in the simulation model. For lowly heritable traits, evaluations obtained by BP under models with up to three loci closely matched the evaluations obtained by BLP for both purebred and crossbred data. For highly heritable traits, models with up to six loci were needed to match the evaluations obtained by BLP.

number of loci / finite locus models / Markov chain Monte Carlo

* Corresponding author: ltotir@iastate.edu

1. INTRODUCTION

Best linear unbiased prediction (BLUP), which can be obtained efficiently by solving Henderson's mixed model equations (HMME) [20], is currently the most widely used method for genetic evaluation. One of the requirements for building HMME is to calculate the inverse of the variance covariance matrix of any random effect in the model. Under additive inheritance, efficient algorithms to calculate the required inverse of the genotypic covariance matrix have been developed for both purebred [18, 19, 27, 28] and crossbred [9, 24] populations. Under non-additive inheritance, algorithms to calculate the required inverse have been investigated as well [21, 30, 35], but these algorithms are not feasible for large inbred populations [6]. This is especially true for crossbred populations [23]. However some traits of interest, for example reproductive or disease resistance traits, are known to have low heritability. Some lowly heritable traits have been shown to exhibit non-additive gene action [5]. Also, the breeding strategies used in several livestock species exploit cross-breeding. Thus, efficient methods for genetic evaluation under non-additive inheritance for purebred and especially for crossbred populations must be developed.

Finite locus models can easily accommodate non-additive inheritance as well as crossbred data. The use of the conditional mean of genotypic values given phenotypes, calculated under the assumption of a finite locus model, has been suggested as an alternative to BLUP [14, 15, 32]. Due to the fact that, conditional on the assumed model being correct, the conditional mean minimizes the mean square error of prediction, and because selection based on the conditional mean maximizes the mean of the selected candidates [2, 13], the conditional mean is also known as the best predictor (BP). Given a finite locus model, the BP can be calculated exactly using Elston-Stewart type algorithms [8], approximated using iterative peeling [34], or estimated using Markov chain Monte Carlo (MCMC) methods [14, 15, 32]. The computational efficiency of these methods is directly related to the number of loci considered in the finite locus model [33]. For Elston-Stewart type algorithms, this relationship is exponential whereas for MCMC methods a linear relationship can be maintained by sampling genotypes one locus at a time.

The exact number of quantitative trait loci (QTL) responsible for the genetic variation of a quantitative trait is not known. However, after performing a meta-analysis on published results from various QTL mapping experiments, Hayes and Goddard estimate that between 50 and 100 loci are segregating in dairy cattle and swine populations [17]. For the large pedigrees encountered in real livestock populations, genetic evaluation by BP using a finite locus model with 50 to 100 loci is computationally unfeasible. Therefore, in this paper,

we investigate the minimum number of loci needed for BP evaluations obtained using a finite locus model to be similar to evaluations obtained by best linear prediction (BLP). Finite locus models with a small number (two through six) of loci (FLMS) were used to obtain evaluations by BP for data sets generated using finite locus models with a large number (about 100) of loci (FLML). These BP evaluations were then compared to BLP evaluations obtained from the same data sets.

2. METHODS

2.1. Notation

Consider a trait determined by N segregating quantitative trait loci (QTL) with two alleles at each locus in a population of n individuals (purebred or crossbred). For convenience, we will use the term reference breed for the purebred or for one of the distinct breed groups in the crossbred population [23]. When only additive and dominance gene action is present, the vector $\boldsymbol{u}$ of genotypic values of the n individuals can be modeled as

$$\begin{aligned} \boldsymbol{u} &= \mathbf{1}\eta + \sum_{i=1}^{N} \boldsymbol{u}_i \\ &= \mathbf{1}\eta + \sum_{i=1}^{N} \boldsymbol{Q}_i \boldsymbol{\delta}_i, \end{aligned} \tag{1}$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones; η is the trait mean in the reference breed; $\boldsymbol{u}_i$ is the $n \times 1$ vector of genotypic values at locus i; $\boldsymbol{Q}_i$ is an $n \times 3$ incidence matrix relating the genotypic values at locus i to the corresponding individuals, with each row of $\boldsymbol{Q}_i$ being one of the vectors [1 0 0], [0 1 0], or [0 0 1]; $\boldsymbol{\delta}_i$ is an 3×1 vector that contains the genotypic effects at locus i: $[a_i \; d_i \; -a_i]'$ [10]. The parameters of this model are: η, the genotypic effects a_i and d_i, and gene frequency p_i, for locus $i = 1, \ldots, N$.

In matrix notation, the vector $\boldsymbol{y}$ of phenotypic values of n individuals can be written as a function of the genotypic values as follows

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{X}$ is the incidence matrix relating the vector $\boldsymbol{\beta}$ of fixed effects to $\boldsymbol{y}$; $\boldsymbol{Z}$ is the incidence matrix relating $\boldsymbol{u}$ to $\boldsymbol{y}$; $\boldsymbol{u}$ is the vector of genotypic values from (1); $\boldsymbol{e}$ is the vector of residuals $\sim \boldsymbol{N}(0, \boldsymbol{I}\sigma_e^2)$.

2.2. Genetic evaluation by BLP

Consider first the situation where $\boldsymbol{u}$ is modeled using a large number of loci each with a small effect. Under such a model, the distribution of genotypic values is approximately multivariate normal. As a result, we can assume that $\boldsymbol{u}$ and $\boldsymbol{y}$ are approximately multivariate normal

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu_u} \\ \boldsymbol{\mu_y} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \boldsymbol{C} \\ \boldsymbol{C}' & \boldsymbol{V} \end{bmatrix} \right), \tag{3}$$

where $\boldsymbol{\mu_u}$ is the vector of genotypic means; $\boldsymbol{\mu_y} = \boldsymbol{X\beta}$; $\boldsymbol{G}$ is the genotypic variance covariance matrix; $\boldsymbol{C} = \boldsymbol{GZ}'$ is the covariance matrix between $\boldsymbol{u}$ and $\boldsymbol{y}$'; $\boldsymbol{V} = \boldsymbol{ZGZ}' + \boldsymbol{I}\sigma_e^2$ is the variance covariance matrix of $\boldsymbol{y}$. Under multivariate normality the conditional mean is also the BLP and can be written as

$$\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y}) = \boldsymbol{\mu_u} + \boldsymbol{CV}^{-1}(\boldsymbol{y} - \boldsymbol{\mu_y}). \tag{4}$$

Note that BLP is a function of the first and second moments of the genotypic values and the phenotypes. The theory for modeling genetic means is well known for both purebred and crossbred populations [4, 7]. The theory for modeling the genetic covariances is also known for both purebred [16, 22] and crossbred [23] populations. However, the covariance theory for crossbred populations is more complex. For example, in a non-inbred, unselected, purebred population, if we ignore linkage and if only additive and dominance gene action are considered, the genetic variance covariance matrix can be written as

$$\boldsymbol{G} = \boldsymbol{A}\sigma_a^2 + \boldsymbol{D}\sigma_d^2, \tag{5}$$

where $\boldsymbol{A}$ is the additive relationship matrix; σ_a^2 is the additive variance; $\boldsymbol{D}$ is the dominance relationship matrix; σ_d^2 is the dominance variance. However, for example, following Fernando [12] in a two breed situation where inbreeding is present the genetic variance covariance matrix becomes

$$\boldsymbol{G} = \sum_{q=1}^{25} \boldsymbol{C}_q \theta_q, \tag{6}$$

where θ_q is the dispersion parameter corresponding to one of 25 breed-specific identity states that specify the breed origin for homologous alleles for a pair of individuals in addition to their identity by descent states [23]; $\boldsymbol{C}_q$ is the matrix of coefficients for θ_q. Recursive formulae are available to compute the elements of $\boldsymbol{C}_q$ [23]. In the absence of inbreeding, the number of dispersion parameters is

reduced from 25 to 12 [23]. Thus, for small pedigrees given known parameters, BLP's can be obtained for both purebred and crossbred populations. For large pedigrees, under non additive inheritance, BLP's cannot be obtained for either purebred or crossbred populations because efficient algorithms to invert $\boldsymbol{G}$ are not available.

2.3. Genetic evaluation by BP

Consider now the situation where $\boldsymbol{u}$ is modeled using a small number of loci. In this situation, BP can be calculated by summing over all possible genotype configurations as follows

$$\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y}) = \mathbf{1}\eta + \sum_{\boldsymbol{g}} \boldsymbol{u}_g \Pr(\boldsymbol{g} \mid \boldsymbol{y}), \tag{7}$$

where $\boldsymbol{u}_g$ is the vector of of genotypic values that corresponds to the genotype configuration $\boldsymbol{g}$, and

$$\Pr(\boldsymbol{g} \mid \boldsymbol{y}) = \frac{\Pr(\boldsymbol{g}, \boldsymbol{y})}{\Pr(\boldsymbol{y})} \propto \Pr(\boldsymbol{y} \mid \boldsymbol{g}) \Pr(\boldsymbol{g}), \tag{8}$$

where $\Pr(\boldsymbol{y} \mid \boldsymbol{g})$ represents the conditional probability of the phenotypes given genotype configuration $\boldsymbol{g}$, and $\Pr(\boldsymbol{g})$ represents the probability of the genotype configuration $\boldsymbol{g}$. Under a finite locus model, efficient methods to calculate these probabilities are available [1, 8]. From equation (7), it can be seen that the key aspect of this type of genetic evaluation is the correct and efficient computation of the sum over all possible genotype configurations. This sum can be calculated exactly using the Elston-Stewart algorithm. This algorithm, however, is computationally feasible only for simple pedigrees and models with up to about three loci. For complex pedigrees and models with more than three loci, MCMC methods hold most promise for the efficient calculation of the desired sum [33]. In this paper, BP evaluations were calculated using the Elston-Stewart algorithm whenever it was computationally feasible. When the use of the Elston-Stewart algorithm was not feasible, BP evaluations were obtained by using an MCMC method called ESIP [11]. ESIP combines the Elston-Stewart algorithm with iterative peeling to generate joint samples from the entire pedigree one locus at a time [11, 33]. In a previous study [33] we have investigated the performance of ESIP when used for genetic evaluation by BP. From the results of that study, it was determined that 50 000 samples from ESIP are sufficient to estimate the BP accurately.

2.4. Parameters for BLP and BP

The first and second moments needed for genetic evaluation by BLP, were calculated from the gene frequencies and genotypic effects of the FLML used to simulate the data. In contrast, for genetic evaluation by BP, the gene frequencies and genotypic effects of the FLMS were chosen, as described below, such that they yielded the same genotypic mean and the same additive and dominance variances as the FLML that was used for simulation. For convenience, we define an N_1 locus model to be "equivalent" to an N_2 locus model ($N_2 > N_1$) if the genotypic means, the additive variances and the dominance variances of the two models are identical.

2.4.1. Parameters for purebred data models

Consider the simple situation when the gene frequency and the additive effect at all loci of a given model are equal. For this case, we discuss below how to assign values to the gene frequencies and the genotypic effects for the FLMS with N_1 loci and the FLML with N_2 loci so that they are "equivalent".

For a simple model of the above type with any even number N of loci, the genotypic mean (η), additive variance (σ_a^2) and dominance variance (σ_d^2) can be written as

$$\begin{aligned} \eta &= 2na(p-q) + 2npqd_1 + 2npqd_2 \\ \sigma_a^2 &= 2npq[a + d_1(q-p)]^2 + 2npq[a + d_2(q-p)]^2 \\ \sigma_d^2 &= n(2pqd_1)^2 + n(2pqd_2)^2, \end{aligned} \tag{9}$$

where $n = \frac{N}{2}$; a is the genotypic effect of one of the homozygotes at the N loci; p is the frequency of one of the two alleles at each of the N loci; $q = 1 - p$; d_1 is the genotypic effect of the heterozygote at half of the N loci and d_2 the genotypic effect of the heterozygote at the other half of the N loci. We simplify further by setting the inbreeding depression ($ID = 2npqd_1 + 2npqd_2$) equal to zero. As a result, d_1 is equal to $-d_2$. Note that in this case, the inbreeding depression is zero while the dominance variance is nonzero. After some algebra, making use of the fact that $q = 1 - p$ and $d_1 = -d_2$, the system

of equations (9) yields

$$p = \frac{\eta + 2na}{4na}$$
$$0 = 16a^4n^4 - a^2(8n^2\eta^2 + 16n^3\sigma_a^2) + \eta^4 + 8n\sigma_d^2\eta^2 + 4n\eta^2\sigma_a^2 \qquad (10)$$
$$d_1 = \frac{\sqrt{\sigma_d^2}}{2p(1-p)\sqrt{2n}}.$$

The second equation in the (10) can be solved for a in terms of n, η, σ_a^2 and σ_d^2. Next, by substituting the value obtained for a in the first equation we can obtain p in terms of n, η, σ_a^2 and σ_d^2, and then by substituting p in the third equation we can obtain d_1 in terms of n, η, σ_a^2 and σ_d^2. Thus, for simple models of this type, the gene frequencies and genotypic effects are completely determined by the genotypic mean, and the additive and dominance variances.

Now consider the two models of interest, a FLMS with N_1 loci, and a FLML with N_2 loci. Under the assumptions described above, the gene frequencies and genotypic effects for each of the two models can be obtained by solving the system of equations given in (10) with $n = \frac{N_1}{2}$ and $n = \frac{N_2}{2}$ respectively, given the assigned values for η, σ_a^2 and σ_d^2. When the number of loci (N) is uneven, at the last locus, the heterozygous genotype is assigned an effect equal to zero ($d_N = 0$).

2.4.2. Parameters for crossbred data models

For the purpose of this paper, crossbred data are simulated by adding k extra loci to the purebred FLML. Thus, crossbred data are simulated with a FLML with $N_2 + k$ loci, where the N_2 loci have the same gene frequency in all breeds and the k loci have different gene frequencies for different breeds. The values for the gene frequencies and genotypic effects for a FLMS with $N_1 + k$ loci are determined, so that it is "equivalent" to the FLML with $N_2 + k$ loci, as follows. First, the FLMS and the FLML are made "equivalent" with respect to N_1 and N_2 loci under a purebred setting. Next, the same gene frequencies and genotypic effects are used for the k extra loci in both models.

2.5. Simulation study

2.5.1. Purebred data

Hypothetical pedigrees. Three hypothetical pedigrees were used to investigate the effect of the number of loci on genetic evaluation by BP. The first

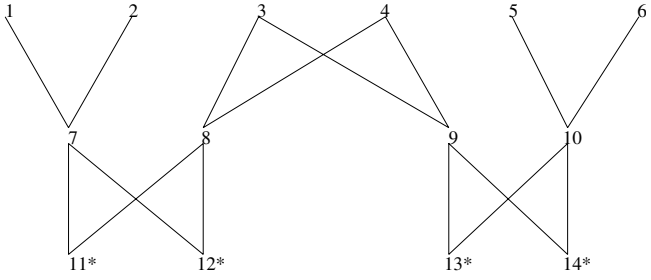

Figure 1. Simple Pedigree. Genetic evaluations were obtained for individuals marked by *.

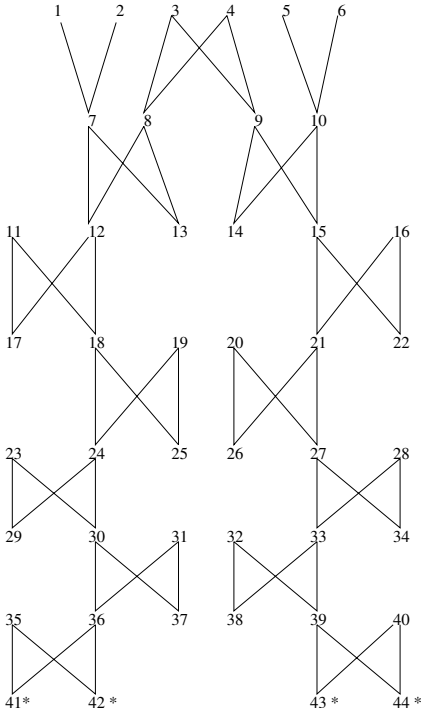

Figure 2. Extended Pedigree. Genetic evaluations were obtained for individuals marked by *.

hypothetical pedigree, shown in Figure 1, has 14 individuals, no loops and will be referred to as the simple pedigree.

The second pedigree, shown in Figure 2, was obtained by extending the first pedigree for five more generations. This pedigree of 44 individuals has eight generations, no loops and will be referred to as the extended pedigree.

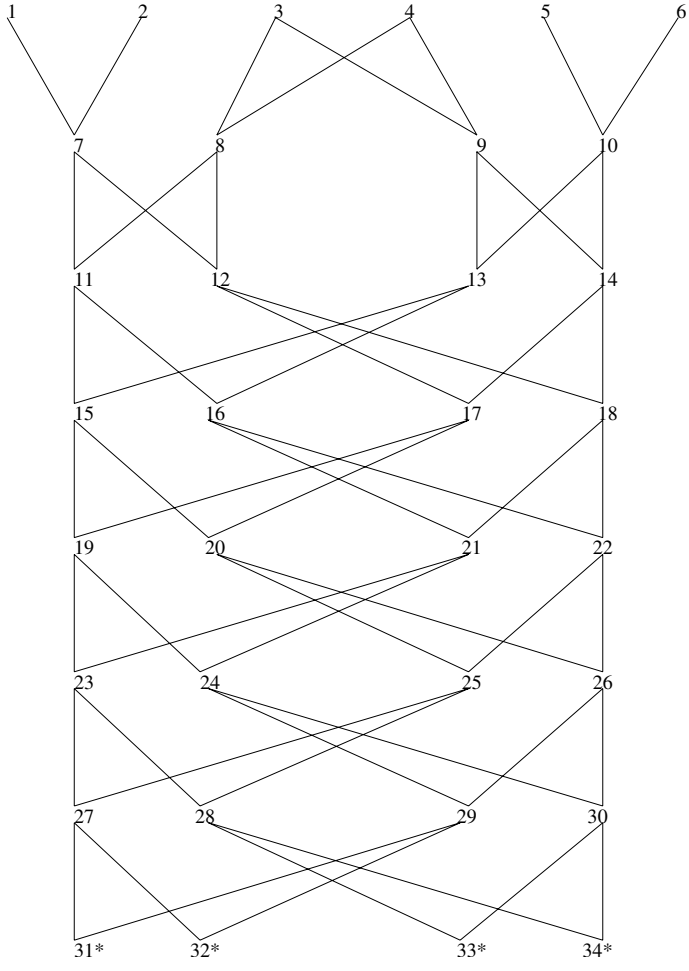

Figure 3. Inbred Pedigree. Genetic evaluations were obtained for individuals marked by *.

The third pedigree, shown in Figure 3, is a highly inbred pedigree with many loops. This pedigree of 34 individuals has eight generations, several loops generated by repeated half sib matings and will be referred to as the inbred pedigree.

Purebred data were simulated using a FLML with 100 loci. At each of the 100 loci, the gene frequency was $p = 0.5$ and the additive effect was $a = 0.2828$. Of the 100 loci, at each of 50, the dominance effect was $d_1 = 0.2828$, and at each of the remaining 50, the dominance effect was $d_2 = -0.2828$. These values yield $\eta = 0$, $\sigma_a^2 = 4$ and $\sigma_d^2 = 2$. Two values were used for the error

Table I. Situations simulated for the purebred case for four different pedigrees. No. missing denotes the number of parents with missing phenotypic information. h_n^2 denotes the narrow sense heritability, and h_b^2 denotes the broad sense heritability.

Situation	Pedigree	No. missing	h_n^2	h_b^2
1	simple	0	0.1	0.15
2	simple	0	0.4	0.6
3	extended	0	0.1	0.15
4	extended	0	0.4	0.6
5	extended	10	0.1	0.15
6	extended	10	0.4	0.6
7	extended	15	0.1	0.15
8	extended	15	0.4	0.6
9	inbred	0	0.1	0.15
10	real	0	0.1	0.15
11	real	0	0.1	0.11

variance: $\sigma_e^2 = 34$ and $\sigma_e^2 = 4$, which combined with the genetic parameters yield two levels of narrow sense heritability: 0.1 and 0.4, with corresponding broad sense heritabilities of 0.15 and 0.6. In order to examine the effect of pedigree structure, missing data, and genetic parameters on genetic evaluations by BP using various FLMS, nine situations were simulated for the hypothetical pedigrees of the purebred case (Tab. I).

The first four situations cover all possible combinations of two heritabilities (0.1 and 0.4) and two types of non inbred pedigrees (simple and extended). This design allows us to examine the main effects of heritability and pedigree size as well as the interactions between these two factors. Situations 3, 4, 5, 6, 7, 8 cover all possible combinations of two heritabilities (0.1 and 0.4) and three patterns of missing data: all individuals have phenotypic data; all individuals in the first two generations have missing data (10 individuals); all sires in the pedigree have missing data (15 individuals). This design allows us to examine the main effects of heritability and missing data as well as the possible interactions between these two factors. Situation 9, which differs from situations 1 and 3 only in the pedigree type, is considered to examine the effect of the presence of inbreeding.

The parameters of the FLMS used to calculate BP's for the data generated according to the nine situations described above, are given in Table II.

Table II. Parameters for the FLMS used to analyze purebred data for situations 1–10. The second column contains the number of loci in the respective FLMS; a denotes the additive effect at all loci; d_1 the dominance effect at half of the loci; d_2 the dominance effect at the other half of the loci; and p the gene frequency at each locus. * Here the dominance effect of the third locus was set to 0.

FLMS	No. loci	a	d_1	d_2	p
FLM(2)	2	2	2	−2	0.5
FLM(3)	3*	1.63	2	−2	0.5
FLM(4)	4	1.4142	1.4142	−1.4142	0.5
FLM(6)	6	1.1547	1.1547	−1.1547	0.5

Table III. Parameters for the FLMS used to analyze purebred data for situation 11. The second column contains the number of loci in the respective FLMS; a denotes the additive effect at all loci; d_1 the dominance effect at half of the loci; d_2 the dominance effect at the other half of the loci; and p the gene frequency at each locus. * Here the dominance effect of the third locus was set to 0.

FLMS	No. loci	a	d_1	d_2	p
FLM(3)	3*	1.5495	1.3685	−1.3685	0.1558

Note that FLM(N_1) denotes the FLMS with N_1 loci, and that each of the FLMS in Table II yields $\eta = 0$, $\sigma_a^2 = 4$ and $\sigma_d^2 = 2$.

Real pedigree. The pedigree structure of a real swine resource population [25] was also used to study the effect of the number of loci on genetic evaluation by BP for a pedigree of moderate size. The pedigree used has a total of 555 animals. Two situations (10 and 11 in Tab. I) were simulated for this pedigree. Situation 10 differs from situations 1, 3 and 9 only in the pedigree used. The data generated according to situation 10 were analyzed using only the FLM(3) with the parameters shown in Table II. For situation 11, a FLML with 100 loci was used to simulate data, each of the 100 loci having a gene frequency of: $p = 0.427$, with an additive effect $a = 0.219$. Of the 100 loci, at each of 50, the dominance effect was $d_1 = 0.104$, and at each of the remaining 50, the dominance effect was $d_2 = -0.104$. These values yield $\eta = -3.2$, $\sigma_a^2 = 2.36$ and $\sigma_d^2 = 0.26$. Combined with an error variance $\sigma_e^2 = 21$ these genetic parameters yield a trait with narrow sense heritability equal to 0.1, and a broad sense heritability equal to 0.11. The data generated according to situation 11 was analyzed using only the FLM(3) with the parameters shown in Table III. These parameters also yield $\eta = -3.2$, $\sigma_a^2 = 2.36$ and $\sigma_d^2 = 0.26$.

Table IV. Situations simulated for the two-breed case for two pedigrees. h_n^2 denotes the narrow sense heritability, and h_b^2 denotes the broad sense heritability.

Situation	Pedigree	h_n^2	h_b^2
1	simple	0.1	0.142
2	simple	0.4	0.57
3	extended	0.1	0.142
4	extended	0.4	0.57

2.5.2. Crossbred data

Two hypothetical pedigrees were used to investigate the effect of the number of loci on genetic evaluations by BP. The first pedigree has the same structure as the one shown in Figure 1. However, individuals 1, 2, 5, 6, 7 and 10 are of breed A, while individuals 3, 4, 8, and 9 are of breed B. Thus, individuals 11, 12, 13 and 14 are crossbred. The second pedigree is also a two-breed pedigree obtained by extending the first pedigree for five more generations. This extension is done in the same way as in the purebred case, but starting with generation three, sires from alternate breeds are used in alternate generations. Thus, an extended two-breed pedigree with 44 individuals and no loops was generated.

Two-breed data were simulated using a FLML with 100 + 1 loci. The gene frequency and genotypic effects for the first 100 loci in both breeds were assigned the same values as the ones used for the purebred case for situations 1−10. For breed A, the extra locus had a gene frequency $p_A = 0.9$, while for breed B the extra locus had a gene frequency $p_B = 0.1$. The genotypic effects for the extra locus in both breeds were: $a = 2$ and $d_1 = 0$. These values yield $\eta_A = 1.6$, $\eta_B = -1.6$, $\sigma^2_{a_A} = \sigma^2_{a_B} = 4.72$ and $\sigma^2_{d_A} = \sigma^2_{d_B} = 2$. Two values were used for the error variance: $\sigma^2_e = 5.08$ or $\sigma^2_e = 40.48$, which combined with the genetic parameters yield two levels of narrow sense heritability: 0.1 and 0.4 with corresponding broad sense heritabilities of 0.142 and 0.57. In order to examine the effect of pedigree structure and genetic parameters on genetic evaluations by BP using various FLMS, four situations were simulated for the two-breed case (Tab. IV).

No missing data were present in these four situations. The design of the simulation allows us to examine the main effects of heritability and pedigree size as well as the interactions between these two factors. Also, it allows us to compare the effect of the number of loci on genetic evaluations by BP in crossbred *versus* purebred situations.

In the following, FLM(N_1, k) denotes the FLMS with $N_1 + k$ loci, where N_1 are the loci that have the same gene frequencies in both breeds and k are the loci that have different gene frequencies in the two breeds. For the BP analysis under the crossbred model FLM(N_1,k), the gene frequencies and genotypic effects for the N_1 loci are identical to those from the BP analysis under the purebred model FLM(N_1) (Tab. II) used for situations 1−10. For the extra $k = 1$ locus, the gene frequencies and genotypic effects used in the simulation were also used in the analysis.

2.5.3. Comparison between BLP and BP evaluations

For each purebred and crossbred situation considered, 100 replicates of the pedigree phenotypes were generated, following the missing pattern of each situation. However, for purebred situations 1−9 and all crossbred situations, the four individuals in the last generation (see Figs. 1, 2, 3) were always assumed to have missing phenotypic data. Genetic evaluations by BLP and BP were then calculated for these four individuals. For situations 10 and 11, only one of the terminal animals was assumed to have missing phenotypic data. Genetic evaluations by BLP and BP were computed for this animal. For each data set, BP evaluations were obtained under one or more FLMS. For traits with low heritability, BP evaluations were obtained under FLM(2), FLM(3) and FLM(4) for purebred data and FLM(2,1) for crossbred data. For traits with high heritability, BP evaluations were obtained under FLM(2), FLM(3), FLM(4) and FLM(6) for purebred data and FLM(2,1), FLM(3,1) and FLM(4,1) for crossbred data. In each replicate, for each individual evaluated, the absolute difference between BLP and BP evaluations was calculated and then scaled by the genetic standard deviation. We will refer to these scaled absolute differences as absolute errors of BP under a FLMS. Thus, except for situations 10 and 11, 400 absolute errors were obtained for each analysis. However, because full sibs with no phenotypic data on themselves or their progeny have the same genetic evaluations, only 200 of these values are unique. For situations 10 and 11, 100 absolute errors were computed for each analysis. Figures 4 and 5 summarize the corresponding 200 values for each of the nine situations of the purebred data case, and each of the four situations of the crossbred data case, in the form of box plots. Figure 4 also summarizes the corresponding 100 values for situations 10 and 11. A box plot is a graphical representation of a distribution [29]. The lower edge of the gray box represents the 25th percentile, the line within the gray box the 50th percentile, and the upper edge the 75th percentile. The lower and the upper whiskers represent the minimum and the maximum.

By visual inspection of the box plots for each situation, we determined the number of loci (in the FLMS) that is adequate for the BP evaluation to closely match the BLP evaluation. For the FLMS that were so deemed to have an adequate number of loci, the correlation between the BP evaluation and the BLP evaluation was greater than or equal to 0.995.

By visual inspection of these figures, we can also make statistical inferences about the impact of heritability, pedigree size, and missing data on the number of loci required for the BP evaluation to closely match the BLP evaluation.

3. RESULTS

3.1. Purebred analysis

3.1.1. Hypothetical pedigrees

Figure 4 summarizes the magnitude of the absolute errors of BP under FLM(2) up to FLM(6) for situations 1−9.

The results obtained for the first four situations allow us to assess the effect of heritability and pedigree size on the number of loci needed for the BP evaluation to closely match the BLP evaluation. For a lowly heritable trait modeled with a FLML with 100 loci, FLMS with two to three loci were adequate for BP evaluations to closely match the BLP evaluations (situations 1 and 3 in Fig. 4). For a highly heritable trait modeled with a FLML with 100 loci, FLMS with six loci were adequate (situations 2 and 4 in Fig. 4). The size of the pedigree had no impact on the number of loci needed (situations 1 *versus* 3, and 2 *versus* 4 in Fig. 4).

The results obtained for situations 3, 4, 5, 6, 7 and 8 allow us to assess the effect of heritability and missing data on the number of loci needed for the BP evaluation to closely match the BLP evaluation. For these six situations we observe again that, highly heritable traits (situations 4, 6, and 8 in Fig. 4), need to be evaluated using FLMS with a larger number of loci than lowly heritable traits (situations 3, 5, and 7 in Fig. 4). Missing data had no impact on the number of loci needed (situations 5, 7 *versus* 3, and 6, 8 *versus* 4 in Fig. 4).

When inbreeding is present (situation 9 in Fig. 4), the magnitude of the absolute errors obtained with FLM(2) and FLM(3) was higher than the magnitude reached under these models in situations 1 and 3, where inbreeding is absent. For situation 9, evaluations by BP under FLM(4) closely match evaluations by BLP.

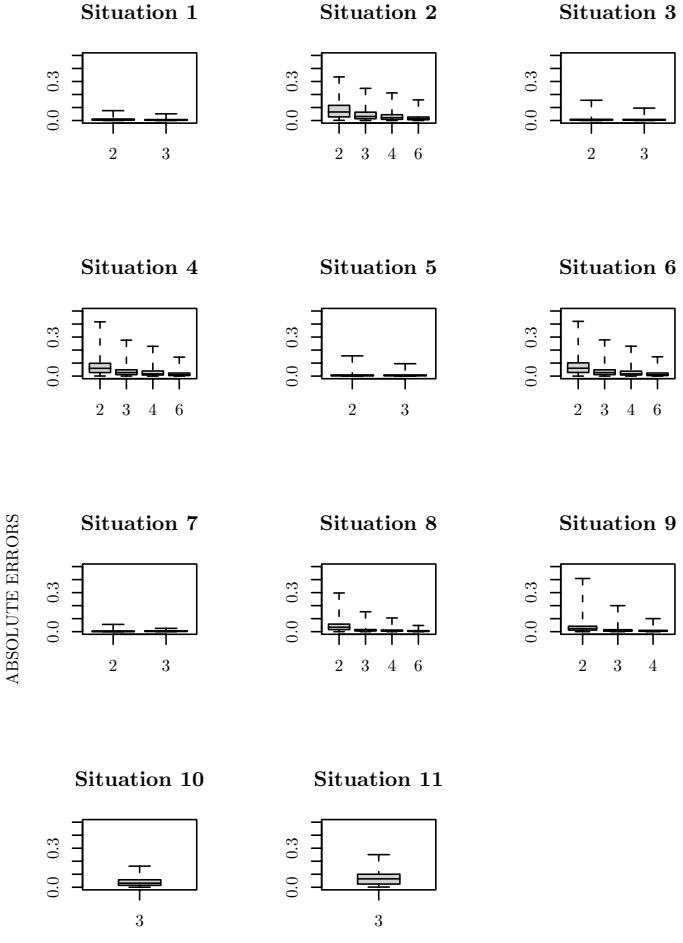

Figure 4. Box plots of 200 (100) values of the absolute errors for BP evaluations under FLM(N), N = 2, 3, 4 or 6 (the X axis of the plots) for situations 1–9 (10, 11) of the purebred case. The units of the Y axis are genetic standard deviations.

3.1.2. Real pedigree

Figure 4 summarizes the magnitude of the absolute errors of BP under FLM(3) for situations 10 and 11. For situation 10, the correlation between the BP evaluation and the BLP evaluation was greater than 0.995 and thus FLM(3) was deemed to have an adequate number of loci. For situation 11, the correlation between the BP evaluation and the BLP evaluation was equal to 0.988 and thus FLM(3) was deemed not to have an adequate number of loci. However, for both situations 10 and 11, BP evaluations were estimated using only 10 000 MCMC samples due to computational constrains. As a result BP

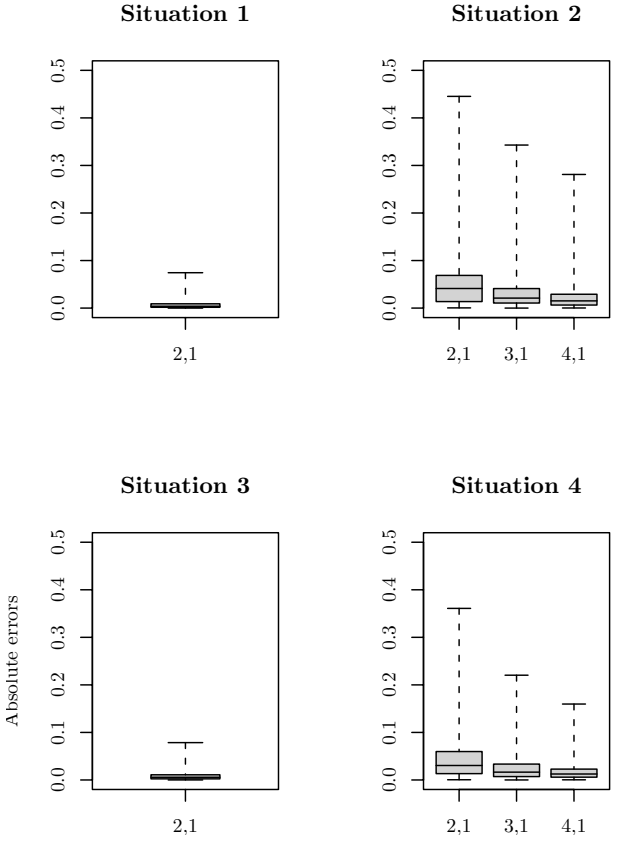

Figure 5. Box plots of 200 values of the absolute errors for BP evaluations under FLM(N, k), N = 2, 3 and 4 and k = 1 (the X axis of the plots) for situations 1−4 of the crossbred case. The units of the Y axis are genetic standard deviations.

evaluations were estimated with less accuracy. For situation 11, FLMS with 4 loci might be needed.

3.2. Crossbred analysis

Figure 5 summarizes the magnitude of the absolute errors of BP under FLM(2,1) up to FLM(4,1) for the four situations of the two-breed case.

For situations 1 and 3 (Fig. 5), evaluations by BP using FLM(2,1) closely match evaluations by BLP. Thus, for a lowly heritable trait modeled with a FLML with 100 + 1 loci, a three locus model was adequate for the BP evaluations to closely match the BLP evaluations. Situations 2 and 4 of the two-breed

case (Fig. 5), correspond to a highly heritable trait (Tab. IV). For these two situations, FLMS with a larger number of loci are needed (Fig. 5). The size of the pedigree had no impact on the number of loci needed (situations 1 *versus* 3, and 2 *versus* 4 in Fig. 5).

4. DISCUSSION

For data simulated using FLML with 100 or 101 loci, evaluations by BP under FLMS with two to six loci matched closely the BLP evaluations. As explained in [33], under dominance inheritance, when inbreeding or crossbreeding is practiced the additive genotypic value of an animal is not a good indicator of the performance of future offspring. Thus, in this paper we did not obtain separate BPs for the additive and the dominance components of the genotypic value. BPs of the genotypic value were calculated instead. The BPs of the genotypic values of future offspring can then be used to select parents.

When MCMC is used to compute the conditional mean (BP), the complete posterior distribution of the genotypic values is available. Thus, the accuracy of the BP evaluations can be represented using the statistic of choice to summarize the posterior distribution (posterior confidence interval, standard deviation, etc.). Finite locus models with small number of loci could also be useful for the problem of parameter estimation in crossbreed populations. Under a linear model, even in a purebred population, large amounts of data are needed to obtain estimates of non-additive effects [3, 12]. In a two breed population with inbreeding, estimates of 5 location and 25 dispersion parameters [23] are needed. Thus, even larger amounts of data would be needed in this case. This, however, might be impractical in livestock populations. By using a finite locus model with a small number of loci, the number of parameters that need to be estimated could be reduced significantly [14, 26, 32]. Thus, parameter estimation in multibreed populations would become practical.

Multiple trait genetic evaluation is warranted only when the traits of interest are correlated. The standard assumption underlying the genetic correlation between traits is pleiotropy [2]. Loci that affect a particular trait combination will have to be modeled separately. For example, suppose loci 1 through 50 affect trait A, loci 51 through 100 affect trait B, and loci 101 through 150 affect both trait A and B. According to the results of this paper, 2 to 6 loci would be needed to model the set of loci (1−50) that affect only trait A, another 2 to 6 for those (51−100) that affect only trait B, and yet another 2 to 6 for those (101−150) that affect both traits A and B. In general, for k traits there can be up to $s = \sum_{i=1}^{k} C_k^i$ distinct subsets of

polygenic loci, where $C_k^i = \frac{k!}{i!(k-i)!}$. When $k = 3, s = 7$ for the trait combinations $\{A, B, C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A\}, \{B\}, \{C\}$. Some of these subsets may be empty or have only a few loci. Thus, the number of loci in the FLMS model may be lower than the maximum of $s \times 6$. MCMC methods such as the reversible jump algorithm [31] can be used to determine the minimum number of loci needed in the FLMS for each of the s subsets. Further research in MCMC methods is needed to make this approach feasible in large livestock pedigrees.

ACKNOWLEDGEMENTS

This journal paper of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 6587, was supported by Hatch Act and State of Iowa funds. This work was partially funded by award No. 2002-35205-1156 of the National Research Initiative Competitive Grants Program of the USDA.

REFERENCES

[1] Bonney G.E., On the statistical determination of major gene mechanisms in continuous human traits: regressive models, Am. J. Med. Genet. 18 (1984) 731–749.

[2] Bulmer M.G., The mathematical theory of quantitative genetics, Clarendon Press, Oxford, 1980.

[3] Chang H.L., Studies on estimation of genetic variances under non-additive gene action, Ph.D. thesis, University of Illinois at Urbana-Champaign, 1988.

[4] Chevalet C., Gillois M., Inbreeding depression and heterosis: Expected means and variances among inbred lines and their crosses, Ann. Genet. Sel. Anim. 10 (1978) 73–98.

[5] Culbertson M.S., Mabry J.W., Misztal I., Gengler N., Bertrand J.K., Varona L., Estimation of dominance variance in purebred yorkshire swine, J. Anim. Sci. 76 (1998) 448–451.

[6] DeBoer I.J.M., Hoeschele I., Genetic evaluation methods for populations with dominance and inbreeding, Theor. Appl. Genet. 86 (1993) 245–258.

[7] Dickerson G.E., Inbreeding and heterosis in animals, in: Anim. Breed. Genet. Symp. in Honor of Dr. J.L. Lush, Champaign, IL, Amer. Soc. Anim. Sci. and Amer. Dairy Sci. Assoc., 1973, pp. 54–77.

[8] Elston R.C., Stewart J., A general model for the genetic analysis of pedigree data, Hum. Hered. 21 (1971) 523–542.

[9] Elzo M.A., Recursive procedures to compute the inverse of the multiple trait additive genetic covariance matrix in inbred and noninbred multibreed populations, J. Anim. Sci. 68 (1990) 1215–1228.

[10] Falconer D.S., Mackay T.F.C., Introduction to quantitative genetics, fourth edition, Longman Inc., New York, 1996.
[11] Fernandez S.A., Fernando R.L., Gulbrandtsen B., Totir L.R., Carriquiry A.L., Sampling genotypes in large pedigrees with loops, Genet. Sel. Evol. 33 (2001) 337–367.
[12] Fernando R.L., Theory for analysis of multi-breed data, in: Proceedings for the Seventh Genetic Prediction Workshop, Kansas City, MO, USA, 3-4 December 1999, pp. 1–16.
[13] Fernando R.L., Gianola D., Optimal properties of the conditional mean as a selection criterion, Theor. Appl. Genet. 72 (1986) 822–825.
[14] Fernando R.L., Grossman M., Genetic evaluation in crossbred populations, in: Proc. Forty-Fifth Annu. Natl. Breeders Roundtable, Poult. Breeders Am. and US Poult. Egg Assoc., Tucker, GA, 1996, pp. 19–28.
[15] Goddard M.E., Gene based models for genetic evaluation - an alternative to blup?, in: Proc. 6th World Cong. Genet. Appl. Livest. Prod., 11-16 January 1998, University of New England, Armidale, Australia, 26 (1998) pp. 33–36.
[16] Harris D.L., Genotypic covariances between inbred relatives, Genetics 50 (1964) 1319–1348.
[17] Hayes B., Goddard M.E., The distribution of the effects of genes affecting quantitative traits in livestock, Genet. Sel. Evol. 33 (2001) 209–229.
[18] Henderson C.R., Sire evaluation and genetic trends, in: Anim. Breed. Genet. Symp. in Honor of Dr. J.L. Lush, Champaign, IL, Amer. Soc. Anim. Sci. and Amer. Dairy Sci. Assoc., 1973, pp. 10–41.
[19] Henderson C.R., A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values, Biometrics 32 (1976) 69–83.
[20] Henderson C.R., Applications of linear models in animal breeding, University of Guelph, Guelph, Ontario, Canada, 1984.
[21] Hoeschele I., VanRaden P.M., Rapid inversion of dominance relationship matrices for noninbred populations by including sire by dam subclass effects, J. Dairy Sci. 74 (1991) 557–569.
[22] Jacquard A., The genetic structure of populations, Springer-Verlag, Germany, 1974.
[23] Lo L.L., Fernando R.L., Cantet R.J.C., Grossman M., Theory of modelling means and covariances in a two-breed population with dominance, Theor. Appl. Genet. 90 (1995) 49–62.
[24] Lo L.L., Fernando R.L., Grossman M., Covariance between relatives in multibreed populations: Additive model, Theor. Appl. Genet. 87 (1993) 423–430.
[25] Malek M., Dekkers J.C.M., Lee H.K., Baas T.J., Rothschild M.F., A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. I. Growth and body composition, Mamm. Genome 12 (2001) 630–636.
[26] Pong-Wong R., Shaw F., Wooliams J.A., Estimation of dominance variation using a finite-locus model, in: Proc. 6th World Cong. Genet. Appl. Livest. Prod., 11-16 January 1998, University of New England, Armidale, Australia, 26 (1998) pp. 41–44.

[27] Quaas R.L., Additive genetic model with groups and relationships, J. Dairy Sci. 71 (1988) 1338–1345.
[28] Quaas R.L., Anderson R.D., Gilmour A.R., BLUP school handbook; use of mixed models for prediction and estimation of (co)variance components, Animal Breeding and Genetics Unit, University of New England, Australia, 1984.
[29] Ramsey F.L., Schafer D.W., The statistical sleuth: a course in methods of data analysis, first edition, Duxbury Press, 1997.
[30] Smith S., Mäki-Tanila A., Genotypic covariance matrices and their inverses for models allowing dominance and inbreeding, Genet. Sel. Evol. 22 (1990) 65–91.
[31] Sorensen D., Gianola D., Likelihood, bayesian and MCMC methods in quantitative genetics, Springer-Verlag, 2002.
[32] Stricker C., Fernando R.L., Some theoretical aspects of finite locus models, in: Proc. 6th World Cong. Genet. Appl. Livest. Prod., 11-16 January 1998, University of New England, Armidale, Australia, 26 (1998) pp. 25–32.
[33] Totir L.R., Fernando R.L., Dekkers J.C.M., Fernandez S.A., Guldbrandtsen B., A comparison of alternative methods to compute conditional genotype probabilities in finite locus models, Genet. Sel. Evol. 35 (2003) 585–604.
[34] van Arendonk J.A.M., Smith C., Kennedy B.W., Method to estimate genotype probabilities at individual loci farm livestock, Theor. Appl. Genet. 78 (1989) 735–740.
[35] VanRaden P., Hoeschele I., Rapid inversion of additive by additive relationship matrices by including sire-dam combination effects, J. Dairy Sci. 74 (1991) 570–579.