Genet. Res., Camb. (2009), **91**, *pp.* 193–207. © Cambridge University Press 2009
doi:10.1017/S0016672309000160 Printed in the United Kingdom

Bayesian prediction of breeding values by accounting for genotype-by-environment interaction in self-pollinating crops

A. M. BAUER[1]*, F. HOTI[2,4], T. C. REETZ[1], W.-D. SCHUH[3], J. LÉON[1] AND M. J. SILLANPÄÄ[4]

[1] *Institute of Crop Science and Resource Conservation, University of Bonn, D-53115 Bonn, Germany*
[2] *Department of Vaccination and Immune Protection, National Institute for Health and Welfare, FIN-00300 Helsinki, Finland*
[3] *Institute of Geodesy and Geoinformation, University of Bonn, D-53115 Bonn, Germany*
[4] *Department of Mathematics and Statistics, Rolf Nevanlinna Institute, University of Helsinki, FIN-00014 Helsinki, Finland*

(*Received 2 September 2008 and in revised form 22 February and 9 April 2009*)

Summary

In self-pollinating populations, individuals are characterized by a high degree of inbreeding. Additionally, phenotypic observations are highly influenced by genotype-by-environment interaction effects. Usually, Bayesian approaches to predict breeding values (in self-pollinating crops) omit genotype-by-environment interactions in the statistical model, which may result in biased estimates. In our study, a Bayesian Gibbs sampling algorithm was developed that is adapted to the high degree of inbreeding in self-pollinated crops and accounts for interaction effects between genotype and environment. As related lines are supposed to show similar genotype-by-environment interaction effects, an extended genetic relationship matrix is included in the Bayesian model. Additionally, since the coefficient matrix C in the mixed model equations can be characterized by rank deficiencies, the pseudoinverse of C was calculated by using the nullspace, which resulted in a faster computation time. In this study, field data of spring barley lines and data of a 'virtual' parental population of self-pollinating crops, generated by computer simulation, were used. For comparison, additional breeding values were predicted by a frequentist approach. In general, standard Bayesian Gibbs sampling and a frequentist approach resulted in similar estimates if heritability of the regarded trait was high. For low heritable traits, the modified Bayesian model, accounting for relatedness between lines in genotype-by-environment interaction, was superior to the standard model.

1. Introduction

The aim of breeding values is to describe the genetic superiority of individuals and hence the capability of individuals to transmit favourable alleles to their progenies. Today the standard approach in animal breeding is to predict breeding values using the mixed model methodology of Best Linear Unbiased Prediction (BLUP) (Henderson, 1984). When applying BLUP, first genetic parameters (variances) are usually estimated by using restricted maximum likelihood (REML) (Patterson & Thompson, 1971). In plant breeding, the prediction of breeding values was considered almost exclusively in research studies (for a review, see Piepho *et al*., 2008). Recently, breeding values have been predicted in self-pollinating crops by accounting for the inbreeding among lines (Bauer *et al*., 2006; Oakey *et al*., 2006; Bauer & Léon, 2008). Selecting by breeding values outperforms commonly used selection strategies, especially if datasets are unbalanced, contain large pedigrees or heritability of the regarded trait is low.

Breeding values are predicted either by a frequentist BLUP approach or by applying Bayesian Gibbs sampling. In general, a frequentist approach provides only the point estimates for breeding values, but separate accuracy estimates based on prediction error variance can be derived afterwards. In a Bayesian framework, by contrast, the whole posterior distribution of the breeding values is estimated conditionally on the

* Corresponding author. e-mail: a.bauer@uni-bonn.de

data and by considering prior information. The Gibbs sampling algorithm, developed by Geman & Geman (1984), led to an increased use of Bayesian methods in quantitative genetics (e.g. Thomas, 1992; Wang *et al.*, 1993; Sorensen *et al.*, 1994). By computing breeding values via Gibbs sampling, accuracy/interval estimates can be obtained since the full marginal posterior distribution of all the parameters of interest is sampled.

In crops, the use of Bayesian Gibbs sampling to predict breeding values is rare. Soria *et al.* (1998) considered Gibbs sampling in a tree-breeding program of Tasmanian bluegum (*Eucalyptus globulus* Labill.) to get inferences about breeding values and genetic parameters. Inferences about major genes and polygenic effects (general and specific combining ability) in a progeny population of Loblolly pine (*Pinus taeda* L.) were obtained by using a parent blocking Gibbs sampler (Zeng *et al.*, 2004). In Scots pine (*Pinus sylvestris* L.), Waldmann & Ericsson (2006) computed a multi-trait individual tree model for a diallele progeny dataset. In a following study, Waldmann *et al.* (2008) developed a fast hybrid Gibbs sampler, which accounted for additive and dominance variances in a mixed model using the same Scots pine dataset as above. Gwaze & Woolliams (2001) estimated a quadrivariate tree model to obtain covariance components for height across two sites and two ages for the Loblolly pine (*P. taeda* L.). In this study, the two sites and two ages were treated as different traits.

Of all the previous quantitative genetics research studies of plant science, Bayesian Gibbs sampling has mainly been applied in forest tree breeding. It seems that the Gibbs sampler has not yet been applied to predicting breeding values in annual crops, especially if they are self-pollinated. Additionally, there is no Gibbs sampling algorithm available for statistical models where genotype-by-environment interactions are included. Genotype-by-environment *interactions* occur if lines react differently to changing environmental conditions resulting often in a different rank order of the lines in each environment (Weber & Wricke, 1990). Therefore, plant breeders are forced to evaluate the lines in several locations and years for estimating the genetic performance. The genotype-by-environment interactions can be analysed for example in a stability analysis (Becker & Léon, 1988) or by an Additive Main Effects Multiplicative Interaction (AMMI) model (Gauch, 1988). Note that the genotype-by-environment interactions being commonly observed in breeding plants and, to a lower extent in animals, are different to the so-called genotype-by-environment *correlations* that occur mainly in animal breeding. Genotype-by-environment correlations can arise when elite animals are kept in a more favourable environment than weaker animals. In this case, the genotype and the corresponding environmental conditions are correlated among each other (Falconer & Mackay, 1996). In plant breeding, usually genotype-by-environment correlation is avoided by adequate experimental field designs due to randomization of environments. As in plant breeding the genotype-by-environment interactions have much more importance, this interaction effect should be accounted for in the Bayesian model to predict breeding values. Up to now, considering interaction effects in Bayesian Gibbs sampling is scarce.

Thus, the objective of our research is to predict Bayesian breeding values of spring barley (*Hordeum vulgare* L.) lines and of a 'virtual' parental population of self-pollinating crops generated by computer simulation. A Gibbs sampling algorithm was developed that (i) is adapted to the high degree of inbreeding in a parental population of self-pollinating crops and (ii) accounts for genotype-by-environment interactions. In addition, standard REML estimates from a frequentist approach were also calculated for comparison.

2. Material and methods

(i) *Simulation*

Following Bauer & Léon (2008), a 'virtual' population of 100 parental inbred lines was generated by Monte Carlo simulation using the Interactive Matrix Language (IML) of SAS software (SAS Institute, 2004). As the Bayesian prediction of breeding values was computationally demanding, a multi-environmental field trial for one year was generated. All lines were cultivated at three locations with three blocks and two replications each. The phenotypic value of a line consisted of a genotypic value, and of random location, genotype-by-location interaction, block and residual effects. The genotypic value of a line was obtained by first generating random additive-genetic effects for each of the 150 loci from a standard normal distribution following the one-locus model and then summing these additive-genetic effects over all 150 loci. In addition, for 10% of the parental lines a random measurement error was computed as systematic noise to make the data simulation more realistic. This heterogeneity leads to an increase of the residual for some genotypes. All simulated effects, except of genotype-by-location interaction and residual effects, were normally distributed with a mean of zero and a standard deviation of one [$N(0, 1)$].

The genotype-by-location interaction effect was simulated by assigning in a first step a normally distributed random number [$N(0, 0{\cdot}5)$] to each possible combination of location and allele (B or b) at a

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

locus as related lines are assumed to have a similar genotype-by-location interaction. Then, for all lines the interaction effects are added over all loci at each location separately. Thus, the more alleles that two lines have in common, the more similar their genotype-by-environment interactions will appear.

In the simulation, all parental lines were measured for a high and a low heritable trait. To obtain traits with a different heritability, the residual effects were generated from a normal distribution with a mean of zero but a varying standard deviation. The high (low) heritable trait was computed from a residual effect with a standard deviation of 17 (56).

The population of parental lines was divided into a base and a progeny population. The base population comprised 30 lines that were randomly crossed with each other to produce 70 progeny lines. The progeny lines were self-pollinated until they reached homozygosity. Hence, all lines in the population were assumed to be homozygous with an inbreeding coefficient of 0·99. For the whole population, pedigree information was available.

The simulation procedure of the parental population was repeated 100 times. The SEEDGEN macro, written by Fan *et al.* (2002), was used to avoid overlapping streams in the generation of random numbers.

(ii) *Field data*

Field data of 82 spring barley (*H. vulgare* L.) lines originating from the German North Rhine–Westphalia core collection (Bauer *et al.*, 2006, 2008) were used. The lines were cultivated in a randomized complete-block design at the Research Station 'Dikopshof' of University of Bonn near to Cologne (Germany) in two different years (2002 and 2003) with three replications each. The lines were measured for the trait 'thousand kernel mass'. Pedigree information was available for all lines.

(iii) *Data analysis*

Breeding values were predicted by a frequentist approach using the software package ASReml 2.0 (Gilmour *et al.*, 2005) as well as by the Bayesian blocked Gibbs sampling algorithm of García-Cortés & Sorensen (1996), which was implemented in Matlab 7 (2007). Both approaches were computed for 100 simulation replicates of the 'virtual' population of parental lines and for the spring barley lines.

Pedigree information was accounted for in the genetic relationship matrix A, which considers the additive-genetic variances and covariances among the lines. Henderson (1976) and Quaas (1976) developed a recursive algorithm to calculate this matrix efficiently. There, the authors assumed that the base population is not inbred. Considering self-pollinating crops, however, usually the base population is highly inbred. Ignoring this degree of inbreeding in the base population would lead to biased breeding values of these lines. Thus, following Bauer & Léon (2008) the coefficient of inbreeding of the lines was accounted for not only for the progenies but also for their parents in the base population by obtaining all diagonal elements of the A-matrix from $1+F_i$ (where F_i equals the coefficient of inbreeding). This is in contrast to Henderson (1976), who did not consider the coefficient of inbreeding in computing the diagonal elements of parental individuals.

(a) *Frequentist approach*

In a frequentist approach, all effects were assumed to be random, so that the X matrix includes only the overall population mean. Therefore, the corresponding linear model of the 'virtual' parental population can be written in matrix notation as

$$y=X\beta+Zu+Ps+Qt+Fh+e,$$

where y is the vector of phenotypic observations; β is the vector of the fixed effect; u is the vector of the random genotypic effect of the lines; s is the vector of the random location effect; t is the vector of the random block effect; h is the vector of the random genotype-by-location interaction effect; and e is the vector of residual effect. X, Z, P, Q and F represent the corresponding design matrices.

Similarly, the statistical model of the spring barley lines follows from:

$$y=X\beta+Zu+Vl+Fh+e,$$

where l is the vector of the random year effect; h is the vector of the random genotype-by-year interaction effect; V and F are design matrices.

For simplification, in the following, the genotype-by-location interaction in the 'virtual' population and the genotype-by-year interaction of the spring barley lines will be summarized as genotype-by-environment interaction.

The resulting covariance structures of the estimated effects are: $\text{Var}(u)=A\sigma_g^2$, $\text{Var}(s)=I\sigma_s^2$, $\text{Var}(t)=I\sigma_t^2$, $\text{Var}(l)=I\sigma_l^2$, $\text{Var}(h)=I\sigma_h^2$, $\text{Var}(e)=I\sigma_e^2=R$ (with $A=$ genetic relationship matrix; $R=$matrix of variances and covariances of residual effects; $I=$identity matrix; $\sigma_g^2=$variance of genotypic effects; $\sigma_s^2=$variance of location effects; $\sigma_t^2=$variance of block effects; $\sigma_l^2=$ variance of year effects; $\sigma_h^2=$variance of genotype-by-environment interaction effects; $\sigma_e^2=$residual variance).

The corresponding Mixed Model Equations (MME) to the statistical model of the 'virtual' population, which are obtained in a frequentist

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

approach, are:

$$\begin{pmatrix} X'X & X'Z & X'P & X'Q & X'F \\ Z'X & Z'Z+A^{-1}\alpha_1 & Z'P & Z'Q & Z'F \\ P'X & P'Z & P'P+I\alpha_2 & P'Q & P'F \\ Q'X & Q'Z & Q'P & Q'Q+I\alpha_3 & Q'F \\ F'X & F'Z & F'P & F'Q & F'F+I\alpha_4 \end{pmatrix} * \begin{pmatrix} \beta \\ u \\ s \\ t \\ h \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \\ P'y \\ Q'y \\ F'y \end{pmatrix},$$

where $\alpha_1=\sigma_e^2/\sigma_g^2$, $\alpha_2=\sigma_e^2/\sigma_s^2$, $\alpha_3=\sigma_e^2/\sigma_t^2$ and $\alpha_4=\sigma_e^2/\sigma_h^2$.

(b) *Bayesian block Gibbs sampling*

In a Bayesian framework, all unknown parameters are sampled from distributions and are therefore treated as random. In this study, the overall mean and the effects for location, block ('virtual' population) and year (field data) were considered in the *X*-matrix. Thus, the statistical model is displayed as

$$y=X\beta+Zu+Fh+e.$$

To compute a Bayesian analysis, prior distributions are assigned to β, u, h, σ_g^2, σ_h^2 and σ_e^2. For β, we assumed an improper prior distribution with $p(\beta)\propto$constant. The prior distributions of u and h were equal to $u|\sigma_g^2\sim N(0,\ A\sigma_g^2)$ and $h|\sigma_h^2\sim N(0,\ I\sigma_h^2)$, respectively. For genotype-by-environment and residual variance, uninformative priors were used. The prior specifications for the variance components were supposed to be scaled inverted chi-square distributions (Gianola & Sorensen, 2002):

$$P(\sigma_i^2|v_i, S_i^2)\propto(\sigma_i^2)^{-(v_i/2+1)}\exp\left(-\frac{v_iS_i^2}{2\sigma_i^2}\right),$$
$$i=u,h,e,$$

where σ_i^2 is the variance component of factor i, v_i is the degree of belief parameter and S_i^2 is the prior value.

In the Bayesian analysis, two approaches were applied that differ in the way the genotype-by-environment interaction is modelled. In the first approach, the variance of the genotype-by-environment interaction was considered similarly to the frequentist approach (REML), where the variance of the genotype-by-environment interaction σ_h^2 is assumed to be independently and identically normally distributed as $h|\sigma_h^2\sim N(0,\ I\sigma_h^2)$. This strategy is referred to as Bayes_ID. In a second approach, we replaced the identity matrix I with an extended relationship matrix $A^{\text{ext}}(=A\otimes I)$ as related lines often show a similar genotype-by-environment interaction (Bayes_A^{ext}).

In the Gibbs sampler, the unknown parameters can be updated and drawn either elementwise (single-site Gibbs sampler) or blockwise (blocked Gibbs sampler). Using the single-site Gibbs sampler, parameters are updated element by element (see e.g. Thomas, 1992; Lin, 1999; Gianola & Sorensen, 2002). As convergence can be very slow, the blocked Gibbs sampler (García-Cortés & Sorensen, 1996) was applied in this study, which drew all parameters as a single block. The disadvantage of this grouped Gibbs sampling is that each iteration needs the inverse of the coefficient matrix *C*, which can slow down the algorithm.

The coefficient matrix *C* of the MME equals

$$C=\begin{pmatrix} X'X & X'Z & X'F \\ Z'X & Z'Z+A^{-1}\alpha_1 & Z'F \\ F'X & F'Z & F'F+I\alpha_4 \end{pmatrix}.$$

The vector of the right-hand side of the MME is $W'y$, with $W=(X\ Z\ F)$. All unknown parameters are summarized in the vector θ. Then according to Gianola & Sorensen (2002) θ can be obtained from

$$\theta=\begin{pmatrix} 0 \\ u \\ h \end{pmatrix}+C^{-1}W'(y-z),$$

where z is a random vector of pseudo-observations with $[z|\mu,\ u,\ h,\ \sigma_e^2]\sim N(X\mu+Zu+Fh,\ I\sigma_e^2)$. Original idea in the García-Cortés & Sorensen (1996) algorithm was to calculate the inverse of the large *C* matrix using iterative methods. However, in the present study, due to relatively small size of *C*, the inverse of *C* was calculated using direct methods. The inverse of *C* is not defined if the *C*-matrix is characterized by rank deficiencies, which occurs if the number of linearly independent rows or columns is smaller than the total number of rows and columns of this matrix. In this case, we calculated the pseudoinverse of *C* using the nullspace U_2 of the *C*-matrix, because this procedure is computationally more efficient than computing simply the pseudoinverse (see Appendix for the derivation). Therefore, θ is calculated from

$$\theta=\begin{pmatrix} 0 \\ u \\ h \end{pmatrix}+(C+U_2U_2')^{-1}*W'(y-z)-U_2U_2'*W'(y-z).$$

The variance components are sampled from

$$\sigma_g^2|\beta,u,h,\sigma_h^2,\sigma_e^2,y\sim\tilde{v}_u\tilde{S}_u^2\chi_{\tilde{v}_u}^2,$$
$$\sigma_h^2|\beta,u,h,\sigma_g^2,\sigma_e^2,y\sim\tilde{v}_h\tilde{S}_h^2\chi_{\tilde{v}_h}^2,\quad\text{and}$$
$$\sigma_e^2|\beta,u,h,\sigma_g^2,\sigma_h^2,y\sim\tilde{v}_e\tilde{S}_e^2\chi_{\tilde{v}_e}^2$$

with $\tilde{v}_u=n+v_u$, $\tilde{v}_h=q+v_h$ and $\tilde{v}_e=N+v_e$ degrees of freedom of a scaled inverted chi-square distribution

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

(n = number of lines; q = number of genotype-by-environment interaction levels; N = number of observations; v_u, v_h, v_e = degree of belief parameter), $\tilde{S}_u^2 = (u'A^{-1}u)/\tilde{v}_u$, $\tilde{S}_h^2 = (h'Ih)/\tilde{v}_h$ and $\tilde{S}_e^2 = [(y - W\theta)'(y - W\theta)]/\tilde{v}_e$. To obtain flat priors, v_u, v_h, v_e were chosen to be -2, and S_u^2, S_h^2, S_e^2 being equal to 0.

The algorithm of the blocked Gibbs sampler is as follows:

1. Initialize the parameters α_1, α_4, σ_g^2, σ_h^2 and σ_e^2. In this study, the starting value for all parameters was equal to 1.
2. Generate u^* from $N(0, A\sigma_g^2)$.
3. Generate h^* from $N(0, I\sigma_h^2)$.
4. Generate z^* from $N(Zu^* + Fh^*, I\sigma_e^2)$.
5. Compute $W'(y - z^*)$.
6. Calculate θ as $\theta = \begin{pmatrix} 0 \\ u^* \\ h^* \end{pmatrix} + C^{-1}W'(y - z^*)$ if C has full rank. Otherwise, if a rank deficiency of C occurs, calculate θ as

$$\theta = \begin{pmatrix} 0 \\ u^* \\ h^* \end{pmatrix} + (C + U_2U_2')^{-1} * W'(y - z^*) - U_2U_2' * W'(y - z^*).$$

7. Calculate $\tilde{S}_u^2$, $\tilde{S}_h^2$ and $\tilde{S}_e^2$.
8. Sample $\tilde{\chi}_i^{-2}$ from $1/\tilde{\chi}_i^{-2}$, where $i = u, h, e$.
9. Compute the variance components from $\tilde{\sigma}_i^2 = \tilde{\chi}_i^{-2}\tilde{S}_i^2$ with $i = u, h, e$.
10. Calculate the variance ratios $\alpha_1 = \tilde{\sigma}_e^2/\tilde{\sigma}_g^2$ and $\alpha_4 = \tilde{\sigma}_e^2/\tilde{\sigma}_h^2$.
11. Update the coefficient matrix C.
12. Repeat steps 1 to 11 until the MCMC chain converges.

As related lines are supposed to have a similar genotype-by-environment interaction, we modified the Gibbs sampler by including an extended genetic relationship matrix A^{ext} in a second run. The extended A^{ext}-matrix was obtained by calculating the Kronecker product of the genetic relationship matrix A with an identity matrix with a size depending on the number of locations (cf. Smith *et al.*, 2001). This extended A^{ext}-matrix then has a block structure. Here, the dataset has to be sorted by location, because it would not otherwise be possible to invert A^{ext}.

Using the extended relationship matrix A^{ext}, instead of the identity matrix, to account for the genotype-by-environment interaction in Gibbs sampling, the coefficient matrix C in the MME is obtained from

$$C = \begin{pmatrix} X'X & X'Z & X'F \\ Z'X & Z'Z + A^{-1}\alpha_1 & Z'F \\ F'X & F'Z & F'F + A^{\text{ext}^{-1}}\alpha_4 \end{pmatrix}.$$

The prior distribution of h equals $h|\sigma_h^2 \sim N(0, A^{\text{ext}}\sigma_h^2)$. So, in the Gibbs sampling, h^* is generated from $N(0, A^{\text{ext}}\sigma_h^2)$. The variance component for genotype-by-environment interaction is sampled from $\sigma_h^2|\beta, u, h, \sigma_g^2, \sigma_e^2, y \sim \tilde{v}_h\tilde{S}_h^2\chi_{\tilde{v}_h}^{-2}$, where $\tilde{S}_h^2 = (h'A^{\text{ext}^{-1}}h)/\tilde{v}_h$.

Accounting for relationship information in the REML variance component estimation of genotype-by-environment interaction variance by using the matrix A^{ext} was not possible due to singularities in the Average Information matrix that is considered in the ASReml program. In Bayesian Gibbs sampling, the problem of singularities in a matrix was solved by accounting for the nullspace of this matrix in the computation of the pseudoinverse.

The Gibbs sampler was run for 50 000 iterations with a 'burn-in' period of 20 000 iterations, although, to save storage capacity, only every 10th sample was considered. The computing time on a Pentium 2·66 GHz dual core processor was a couple of minutes for the frequentist approach and one week for the Bayesian analyses of all simulation replicates.

(iv) *Evaluating results from frequentist approach and Bayesian analyses*

For each simulation replicate and for the spring barley lines, in the Bayesian analyses, point (mean, median and mode) and interval (95% highest posterior density region) estimates of posterior distribution of variance components were calculated using Matlab 7 (2007). Following Hoti *et al.* (2002), a kernel smoothing approach was used to summarize the posterior distribution for mode estimation. Also, the standard deviations of the posterior variance component distributions were derived.

To obtain Bayesian breeding values, point estimates (mean, median and kernel-density-based mode) of the posterior distributions of the estimated genetic line effects were computed. Then, for each analysis (frequentist approach, Bayes_ID and Bayes_A$^{\text{ext}}$) of the 'virtual' population, Spearman's rank correlation coefficient, between estimated breeding values and true genotypic values of the lines, was calculated using the software package SAS 9.1 (SAS Institute, 2004). Spearman's correlation coefficient is derived by first ranking the data and then using the ranks in the formula of Pearson's correlation coefficient. In addition, prediction error variance was derived by computing the variance of the difference between estimated breeding values and true genotypic values.

To summarize the results of the 'virtual' population, the arithmetic means and standard deviations of the REML estimated variance components, Bayesian point and interval estimates, standard deviations of posterior distributions and true (simulated) variance components were computed over all simulation replicates.

The heritability h^2 was calculated for all traits based on true (simulated) variance components (in the 'virtual' population) and on estimated values of variance components given by the frequentist approach, Bayes_ID and Bayes_A^{ext} analyses. To compute the heritability of traits measured in plant breeding trials, one has to account for the fact that in contrast to animals where the heritability usually is based on the individual itself as reference unit, the same (homozygous) plant genotype can be cultivated in replicated field plots in several environments. In the milk production of cows at different days in lactation, a related situation can occur where multiple observations on the same animal were obtained, which give rise to permanent environmental variance. To choose the plant genotype as reference unit for computing the heritability, the fraction of the phenotypic variance being transmitted to the progenies has to be considered. Thus, in the present case, the variances of genotype-by-environment and residual have to be divided by the number of locations, blocks and replications as follows (Hanson, 1963):

$$h^2=\frac{\sigma_g^2}{\sigma_g^2+(\sigma_h^2/j)+(\sigma_e^2/j*k*l)},$$

where σ_g^2 is the variance of genotypic effects; σ_h^2 is the variance of genotype-by-environment interaction effects; σ_e^2 is the residual variance; j is the number of locations; k is the number of blocks; and l is the number of replications within each block and location.

For the Bayesian analyses, first based on the estimated variance components the posterior distributions of heritabilities were computed. Then, the posterior mean, median, mode, standard deviation and 95% highest posterior density interval of the heritabilities were obtained.

3. Results

In this study, genetic effects were predicted by a frequentist approach and two variants of Bayesian methods (Bayes_ID and Bayes_A^{ext}). In Bayesian analyses, the mean, median and mode were calculated as point estimates of the posterior distributions. For comparison, in the 'virtual' population true (simulated) values were also given. The results of the variance component and breeding value estimation of the 'virtual' population and of the spring barley lines will be presented below.

(i) '*Virtual*' *parental population*

In general, with increasing heritability, the estimated variance components correspond to the true values to a greater extent, the standard deviation of posterior distributions is smaller and the posterior distribution is less skewed (Table 1, Fig. 1). In addition, for the genotype-by-environment interaction variance (a small variance component), the posterior distribution is more skewed than for the residual variance (a large variance component) (Fig. 1). In comparing the Bayesian point estimates, except for the case of the Bayes_ID analysis of a low heritable trait, the mean, median and mode were estimated close together and located within the 95% highest posterior density region. If the Bayes_ID method is used for a trait with low heritability, the mean is the only estimate that gives reasonable results (Table 1).

For both traits, the additive-genetic variance is overestimated by the Bayes_A^{ext} method, whereas the Bayes_ID values correspond well to the REML and true (simulated) values. If the heritability of the regarded trait is high, the genotype-by-environment interaction variance is underestimated when the Bayes_A^{ext} analysis is performed, but is estimated accurately with the Bayes_ID and frequentist approaches. With low heritability, erroneous interaction variance components were obtained using the Bayes_ID method. Using the Bayes_A^{ext} or frequentist approaches, however, an estimation of the genotype-by-environment interaction was possible, resulting in overestimated variance components. The estimation of residual variance components yielded slightly overestimated values for Bayesian analyses for both traits (Table 1, Fig. 1). Independently, if the Bayes_ID or Bayes_A^{ext} method is computed, the differences between the estimated and true variance components are higher for additive and genotype-by-environment interaction variances than for the residual variance components.

Heritability was estimated from variance components obtained from REML, Bayes_ID and Bayes_A^{ext}. Similar heritability estimates were found for Bayes_ID and REML, which correspond to the true (simulated) heritability. In contrast, when using Bayes_A^{ext}, slightly overestimated heritability estimates were observed.

To determine how accurate the breeding values were predicted by the frequentist, Bayes_ID and Bayes_A^{ext} approaches, Spearman's rank correlation coefficient between estimated breeding value and true (simulated) genotypic value of the lines and the prediction error variance were calculated for each analysis (Tables 2 and 3). In general, the rank correlation coefficient is higher and the prediction error variance is lower for a high heritable trait than for a trait with low heritability. If we consider a high heritable trait, similar rank correlation coefficients and prediction error variances are obtained for all analyses. By contrast for a low heritable trait, the rank correlation coefficient is maximized and the

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

Table 1. *Posterior mean, median, mode, standard deviation (std) and 95% highest posterior density (HPD) obtained from Bayes_ID and Bayes_A^{ext} methods, estimates from the frequentist approach (REML) and true (simulated) values of variance components for two traits of the 'virtual' population (with additive genetic variance σ_g^2, genotype-by-environment interaction variance σ_h^2, residual variance σ_e^2). In addition, heritability (h^2) estimates are displayed. All estimates were averaged over the 100 simulation replicates*

	Bayes_ID						Bayes_A^{ext}							
	Mean	Median	Mode	Std	HPD 2.5	HPD 97.5	Mean	Median	Mode	Std	HPD 2.5	HPD 97.5	REML	True
High heritable trait														
σ_g^2	77·72	74·87	69·10	2·47	50·83	113·38	88·13	85·31	74·27	2·37	58·88	126·92	75·80	70·74
σ_h^2	37·14	36·16	34·98	0·99	18·39	59·52	21·57	19·08	14·47	2·06	9·51	43·00	38·11	38·03
σ_e^2	290·69	290·40	294·27	0·51	267·55	315·70	302·19	300·58	297·49	0·67	276·53	329·71	288·88	288·98
h^2	0·72	0·72	0·72	0·05	0·61	0·81	0·78	0·78	0·78	0·04	0·69	0·84	0·72	0·71
Low heritable trait														
σ_g^2	73·15	69·50	0·30	9·92	25·60	145·08	116·11	108·09	80·65	6·96	46·85	215·64	73·01	70·74
σ_h^2	25·32	0	0·20	33·56	1·20	104·14	89·86	85·81	77·39	2·92	48·23	151·97	59·24	38·03
σ_e^2	3184·20	3178·10	3101·00	7·92	2944·44	3439·80	3192·50	3190·30	3146·20	5·99	2946·96	3456·46	3136·87	3153·5
h^2	0·26	0·26	0	0·08	0·11	0·43	0·34	0·34	0·34	0·08	0·18	0·50	0·27	0·27

prediction error variance is minimized for the frequentist approach and the Bayes_A^{ext} analysis. The lowest rank correlation and the largest prediction error variance were obtained with the Bayes_ID method.

Considering the standard deviation over the 100 simulation replicates in the 'virtual' population, the standard deviation of the variance components of genotype-by-environment interaction is similar to that of the true genotype-by-environment interaction variance if the Bayes_A^{ext} method was used (Table 4). In contrast, the standard deviation of the variance components of genotype-by-environment interaction obtained by Bayes_ID or a frequentist approach is increased greatly over the simulation replicates for both traits. For a low heritable trait, the Bayes_A^{ext} method and a frequentist approach result in lower standard deviations of Spearman's rank correlation coefficient and prediction error variance over all simulation replicates than Bayes_ID analysis.

(ii) *Spring barley lines*

The genetic parameters of the spring barley lines were predicted by Bayes_ID, Bayes_A^{ext} and a frequentist approach. For all analyses, similar variance components were observed (Table 5). In addition, the trait heritability, the highest posterior density regions and the standard deviation of the posterior distributions of all variance components are in the same range for the Bayes_ID and the Bayes_A^{ext} method. By comparing the point estimates of the posterior distributions, as in the 'virtual' parental population, the mode estimate is smaller than the median that is smaller than the mean (Table 5, Fig. 2).

4. Discussion and conclusions

In this study, a Bayesian model was used to account for genotype-by-environment interaction in two different ways. In the first Bayesian analysis (Bayes_ID), the interaction effect was modelled and treated exactly as in the frequentist approach. However, as related lines are assumed to have a similar genotype-by-environment interaction, we modified the model by including an extended genetic relationship matrix A^{ext}, in a second Bayesian analysis (Bayes_A^{ext}). The objective of the current study was to determine if Bayes_A^{ext} leads to more accurate breeding values than Bayes_ID. For comparison, breeding values were also predicted using a frequentist approach. To estimate the genetic parameters, multi-environmental data of a 'virtual' parental population were considered. To verify the results obtained by Bayesian analyses and a frequentist approach in the 'virtual' population, additionally field data of spring barley lines were used.

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

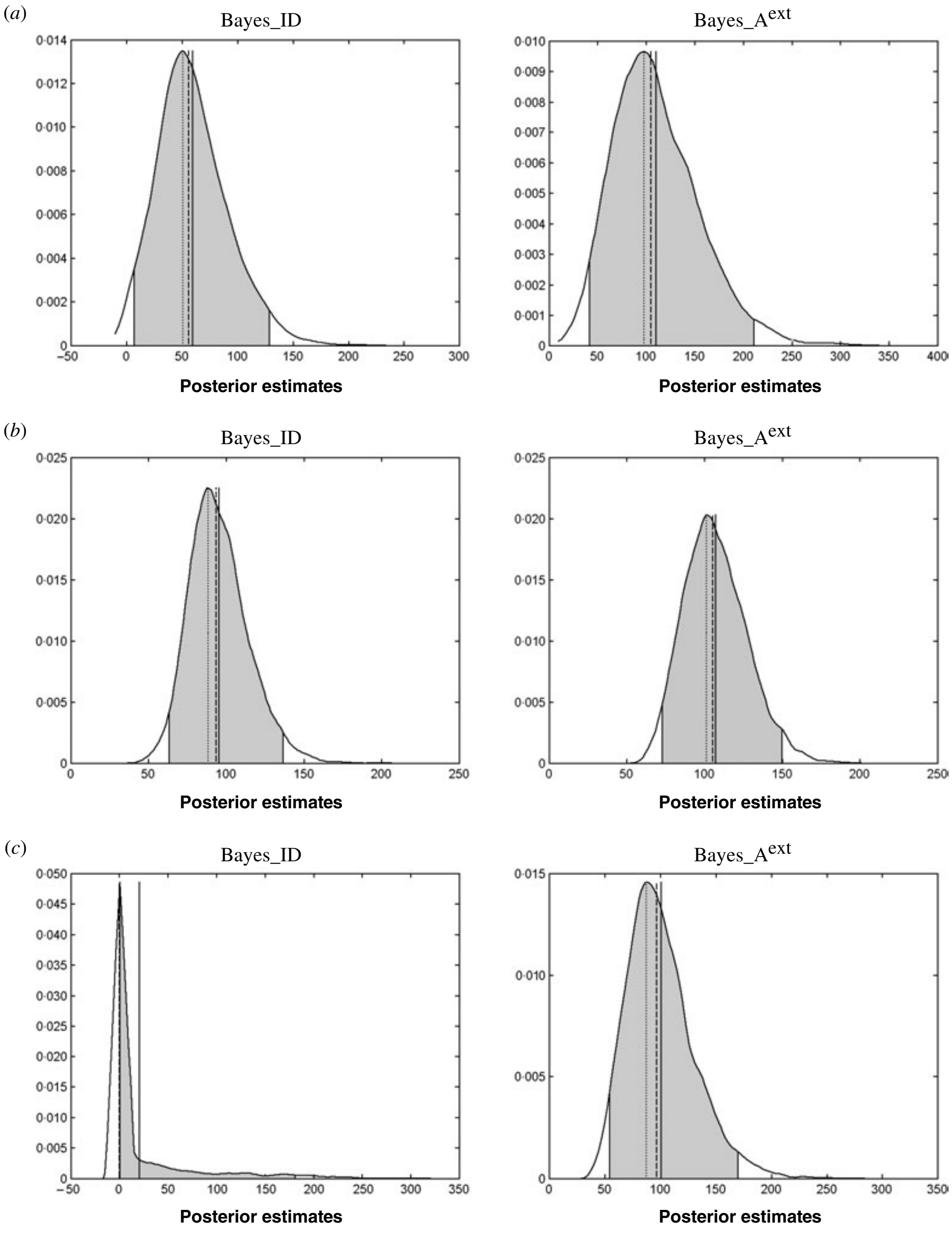

Fig. 1. Continued.

In the frequentist approach using ASReml software (Gilmour *et al.*, 2005), the estimation is divided into two steps. First, the variance components are estimated by REML (Patterson & Thompson, 1971), by applying the Average Information algorithm of Gilmour *et al.* (1995). Next, the estimated REML variance components are used in the MME to predict breeding values. The disadvantage of this strategy is that the uncertainty of the estimated variance components is underestimated because it is not incorporated in the BLUP estimates. By contrast in the Bayesian analyses, it is possible to estimate the variance components and breeding values simultaneously by Gibbs sampling. Thus, the uncertainty of estimated variance components can be accounted for. Getting accurate estimates of the variance components is important because biased estimates can lead to increased prediction errors of breeding values (van Tassell *et al.*, 1995). Due to the mentioned differences between a Bayesian and a frequentist framework, one should have in mind that the methods are not fully comparable, although in this study the general model is the same in Bayesian and frequentist approaches.

In our study, in the 'virtual' parental population, with increasing heritability of the trait, the differences between the estimated variance components and true

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

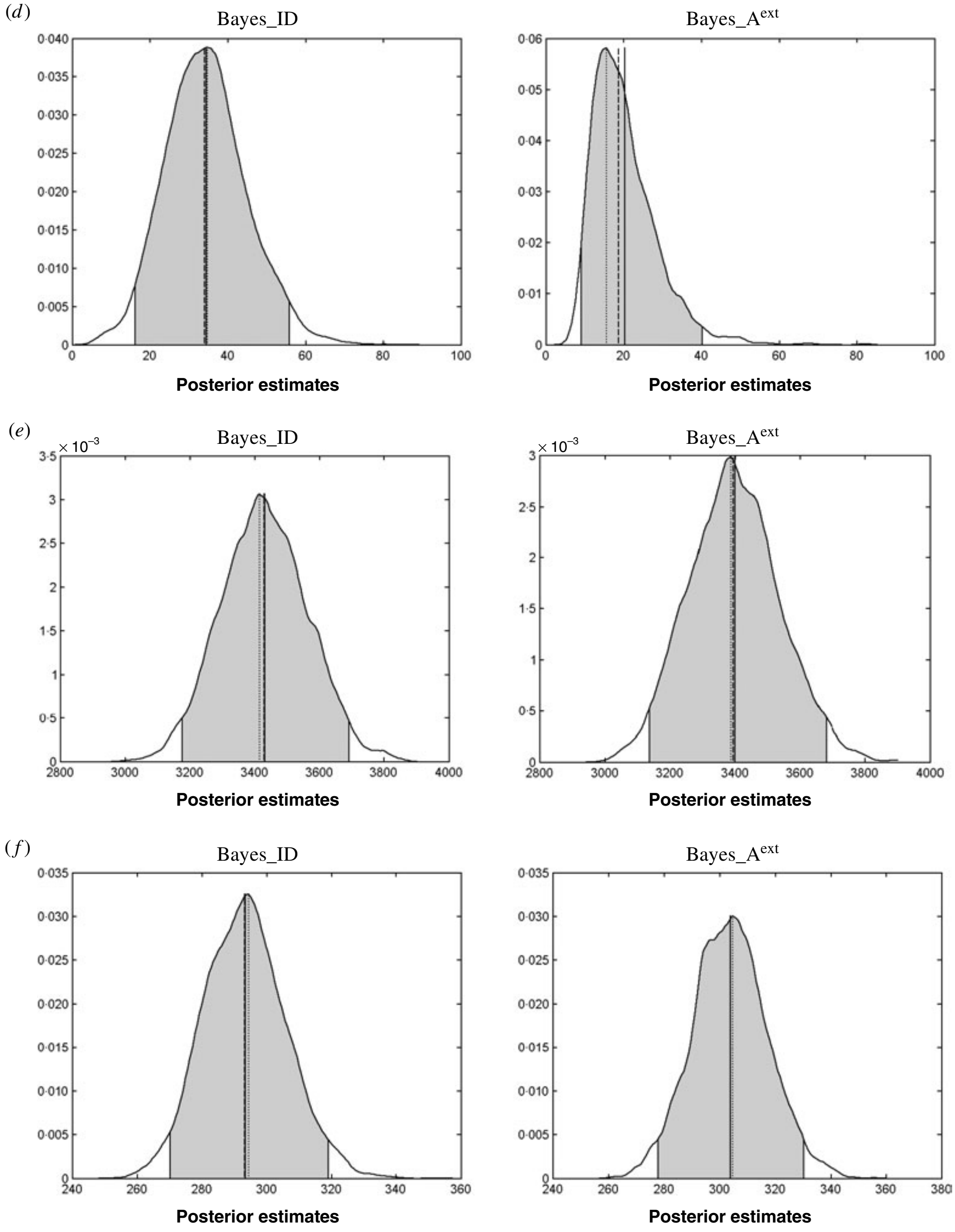

Fig. 1. Frequency distribution of the posterior variance components obtained by Bayes_ID and Bayes_A^{ext} for two traits in the 'virtual' population. Additionally, point estimates and the 95 % highest posterior density regions are given (straight line = posterior mean; dashed line = posterior median; dotted line = posterior mode). In this figure, the results of one simulation replicate are displayed. (*a*) Additive genetic variance of a low heritable trait. (*b*) Additive genetic variance of a high heritable trait. (*c*) Genotype-by-environment interaction variance of a low heritable trait. (*d*) Genotype-by-environment interaction variance of a high heritable trait. (*e*) Residual variance of a low heritable trait. (*f*) Residual variance of a high heritable trait.

(simulated) values decreased, and the standard deviation of posterior distribution was lowered. Phenotypic observations of a high heritable trait are mainly influenced by genetic effects, rather than environmental effects, which leads to an increased prediction accuracy. For additive-genetic and genotype-by-environment interaction variance components, slightly skewed posterior distributions were obtained (Figs 1 and 2). This skewness occurs especially for small variance components (van Tassell *et al.*, 1995), which could be due to the fact that variance components are allowed to only take positive values (Hazelton & Gurrin, 2003). Skewed posterior distributions can result in biased point estimates, such as the mean,

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

Table 2. *Spearman's rank correlation coefficient between true genotypic values and point estimates of breeding values obtained by a frequentist approach and Bayesian analyses of a high and a low heritable trait in the 'virtual' population. The rank correlation coefficients were averaged over all simulation replicates*

	High heritable trait	Low heritable trait
Frequentist	0·90	0·67
Bayes_ID		
Mean	0·90	0·65
Median	0·90	0·64
Mode	0·90	0·64
Bayes_A^{ext}		
Mean	0·90	0·67
Median	0·90	0·66
Mode	0·90	0·66

Table 3. *Prediction error variance of breeding values obtained by a frequentist approach and Bayesian point estimates of a high and a low heritable trait in the 'virtual' population. The prediction error variances were averaged over all simulation replicates*

	High heritable trait	Low heritable trait
Frequentist	23·02	74·76
Bayes_ID		
Mean	23·05	78·79
Median	23·05	81·05
Mode	23·12	82·04
Bayes_A^{ext}		
Mean	23·42	74·66
Median	23·42	74·98
Mode	23·50	75·65

median and mode (Zeger & Karim, 1991; Burton *et al.*, 1999), which differ highly among each other (Hazelton & Gurrin, 2003). In our study, the differences between the estimated variance components and true (simulated) values were higher for additive-genetic and genotype-by-environment interactions than for residual variance (Table 1, Fig. 1). This is because of the skewed additive-genetic and genotype-by-environment interaction posterior distributions, whereas the residual posterior variance was closer to a normal distribution. Waldmann & Ericsson (2006) and Waldmann *et al.* (2008) also stated that the Scots pine (*P. sylvestris* L.) estimates of genetic variance components were more biased due to the posterior distributions being more skewed than the residual variance. In our field data example of spring barley lines, similar variance components were obtained by all prediction strategies (Table 5). For all variance components, we found similar values for the mean, median and mode (Tables 1 and 5). Van Tassell *et al.* (1995) stated that the mean is more appropriate for estimating the variance components than the mode. Additionally, with decreasing heritability of the regarded traits, the differences between the estimates increased, which is also in accordance with Waldmann & Ericsson (2006).

For estimating the mode of a posterior distribution, the mode is usually computed based on a histogram of the MCMC samples. However, this approach depends on the bin width and the sideway shift of the bin grids of the histogram leading to biased mode estimates. Hoti *et al.* (2002) introduced a kernel density estimation to smooth the shape of the posterior distribution. The use of kernel smoothing can improve the localization of the mode estimate significantly.

Heritability estimates provided by REML and Bayes_ID in the 'virtual' population were found to be quite similar to heritability estimates from simulated data (Table 1). In contrast, heritability estimates of the Bayes_A^{ext} variance components were slightly overestimated, since, in this analysis, the additive-genetic variance was positively biased. This could be due to the fact that the genetic relationship matrix was accounted for twice in the Bayes_A^{ext} model. Gwaze & Woolliams (2001) also found a larger heritability when computed with a Bayesian analysis, rather than a REML analysis. In contrast, in the spring barley population Bayes_ID method resulted in a larger heritability estimate than REML, but the lowest heritability estimate was found using Bayes_A^{ext} analysis.

In self-pollinated crops, breeding values are estimates of the true genotypic value of the lines. To predict breeding values having BLUP characteristics, a frequentist approach or a Bayesian method can be used. All strategies should maximize the correlation between true and estimated breeding values and minimize the prediction error variance. Thus, in this study for each analysis, Spearman's rank correlation coefficient between breeding value and true genotypic value (Table 2) and the prediction error variance of estimated breeding values (Table 3) were computed. The higher the rank correlation and the lower the prediction error variance, the more accurate the corresponding prediction method. In general, with decreasing heritability the rank correlation between estimated breeding value and true genotypic value will be lower because of the larger environmental influence on the phenotype, and hence the prediction error variance will be higher. Especially for a low heritable

Table 4. *Standard deviation over simulation replicates of posterior mean, median, mode, standard deviation (post. std) and 95% highest posterior density (HPD) obtained from Bayes_ID and Bayes_A^{ext} methods, and estimates from the frequentist approach (REML) for variance component estimates, Spearman's rank correlation coefficient and prediction error variance of two traits having a high and a low heritability h^2 in the 'virtual' population (with additive genetic variance σ_g^2, genotype-by-environment interaction variance σ_h^2, residual variance σ_e^2)*

	Variance components						Spearman's rank correlation		Prediction error variance	
	High heritability h^2			Low heritability h^2			High h^2	Low h^2	High h^2	Low h^2
	$\sigma_g{}^2$	$\sigma_h{}^2$	$\sigma_e{}^2$	$\sigma_g{}^2$	$\sigma_h{}^2$	$\sigma_e{}^2$				
Frequentist	15·29	10·40	11·05	31·63	61·07	131·39	0·03	0·08	3·43	15·33
Bayes_ID										
Mean	15·71	11·77	11·23	38·97	44·68	135·27	0·03	0·14	3·42	17·84
Median	15·49	11·85	11·24	39·02	43·48	135·19	0·03	0·16	3·43	20·85
Mode	15·24	11·98	11·21	37·15	27·77	132·04	0·03	0·15	3·41	20·62
Post. std	2·47	0·99	0·51	9·92	33·55	7·92	–	–	–	–
HPD 2.5	11·77	9·68	10·33	23·32	9·90	126·84	–	–	–	–
HPD 97.5	21·09	13·52	12·19	58·46	115·89	143·97	–	–	–	–
Bayes_A^{ext}										
Mean	14·96	4·51	11·84	33·80	8·37	131·71	0·03	0·08	3·59	15·46
Median	14·77	4·16	11·87	33·84	7·90	131·34	0·03	0·08	3·60	16·07
Mode	14·03	3·38	12·34	31·24	7·85	130·01	0·03	0·08	3·62	15·53
Post. std	2·37	2·06	0·67	6·96	2·92	5·99	–	–	–	–
HPD 2.5	11·07	1·71	11·00	21·82	4·25	121·61	–	–	–	–
HPD 97.5	20·04	9·38	12·79	47·24	15·21	143·32	–	–	–	–
True	16·29	3·23	10·45	16·29	3·23	106·75	–	–	–	–

Table 5. *Posterior mean, median, mode, standard deviation and 95% highest posterior density (HPD) obtained from Bayes_ID and Bayes_A^{ext} methods, and estimates from the frequentist approach (REML) for the trait 'thousand kernel mass' of the spring barley lines (with additive genetic variance σ_g^2, genotype-by-environment interaction variance σ_h^2, residual variance σ_e^2). In addition, heritability (h^2) estimates are displayed*

	Bayes_ID						Bayes_A^{ext}						
	Mean	Median	Mode	Std	HPD 2.5	HPD 97.5	Mean	Median	Mode	Std	HPD 2.5	HPD 97.5	REML
σ_g^2	7·73	7·47	7·24	2·09	4·11	12·28	7·93	7·72	7·29	2·14	4·40	12·63	7·18
σ_h^2	1·87	1·72	1·63	1·60	0	5·34	3·86	3·62	3·32	1·54	1·63	7·45	2·75
σ_e^2	15·37	15·29	15·13	1·33	12·92	18·19	14·92	14·87	14·76	1·20	12·74	17·36	14·67
h^2	0·69	0·69	0·68				0·64	0·64	0·64				0·65

trait, the prediction strategy Bayes_A^{ext} is superior to Bayes_ID (Tables 2 and 3). Predicting breeding values by using the Bayes_A^{ext} method or a frequentist approach, the estimated breeding values correspond to the true genotypic value to a greater extent than using the Bayes_ID method. Thus, it seems to be important to account for the relationship information between lines not only in predicting the genetic line effect but also in computing genotype-by-environment interactions (Bayes_A^{ext}), if a Bayesian prediction strategy is applied considering genotype-by-environment interactions in the statistical model. In contrast, in a frequentist approach (REML), the integration of relationship information in the calculation of genotype-by-environment interaction variance resulted in singularities in the Average Information matrix. In animal breeding, Schenkel *et al.* (2002) did not find any differences between the rank correlations of a Bayesian or a frequentist approach to simulated breeding values. Also, Robinson (1991) and Harville & Carriquirry (1992) stated that the differences between breeding values predicted by a frequentist or a Bayesian approach are minimal. The superiority of Bayes_A^{ext} analysis for a low heritable trait is supported by considering the standard deviations of estimated genotype-by-environment interaction

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

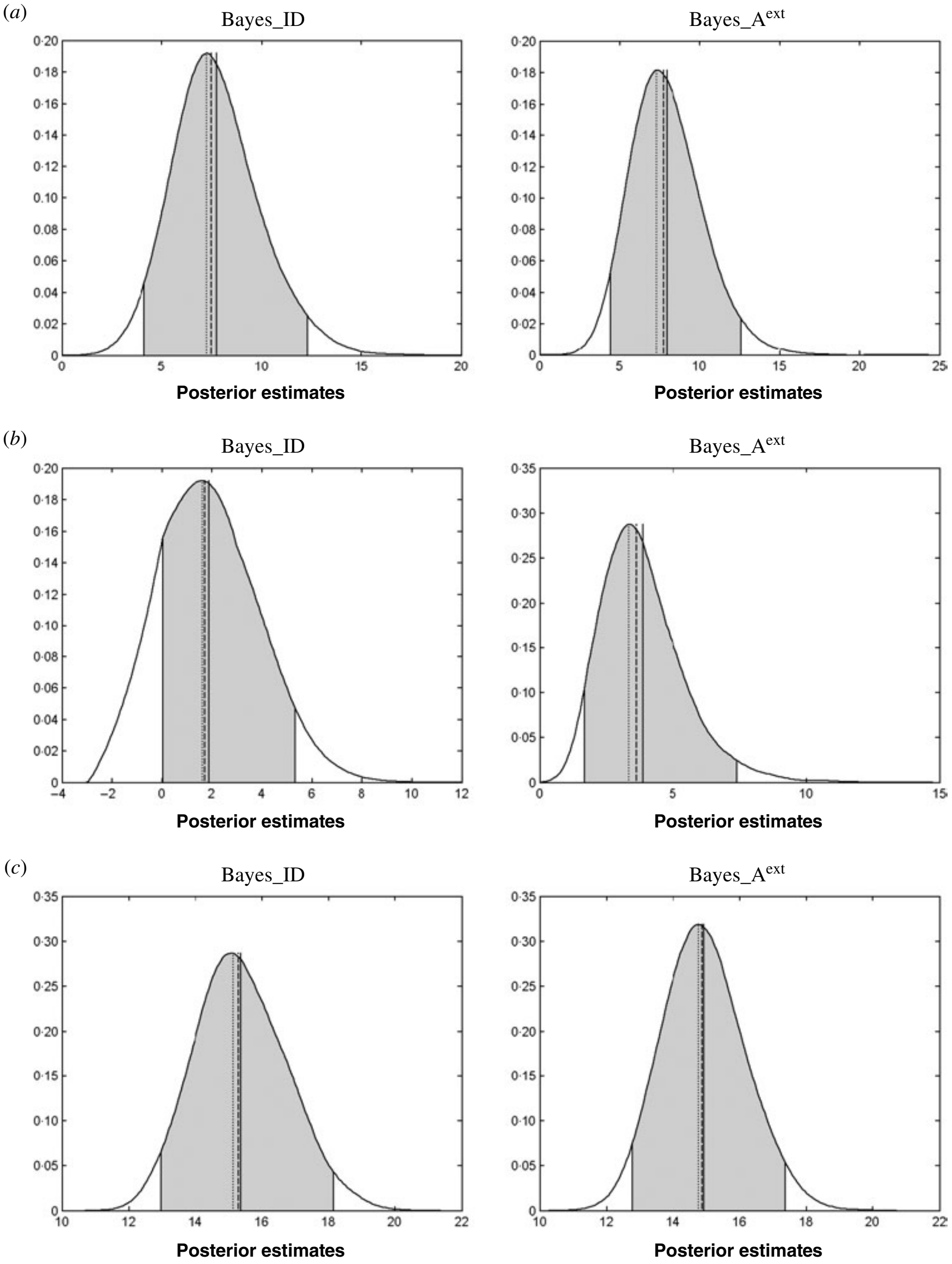

Fig. 2. Frequency distribution of the posterior variance components obtained by Bayes_ID and Bayes_A^{ext} for the trait 'thousand kernel mass' of the spring barley lines. Additionally, point estimates and the 95% highest posterior density regions are given (straight line = posterior mean; dashed line = posterior median; dotted line = posterior mode). (*a*) Additive genetic variance. (*b*) Genotype-by-environment interaction variance. (*c*) Residual variance.

variances, Spearman's rank correlation coefficients and prediction error variances over all simulation replicates (Table 4). Especially for a low heritable trait, a lower standard deviation over the simulation replicates was obtained by the Bayes_A^{ext} method being similar to that of true (simulated) values than with Bayes_ID analysis.

If genotype and environment are correlated among each other as it often occurs in animal breeding, heterogeneity of the residual variance can be found. In such situations, for each environment an own residual could be considered in the model (Fernando *et al.*, 1984) to avoid occurrence of erroneous genotype-by-environment interaction variance due to heterogeneity of variances. In plant breeding, however, due to the fact that the lines are completely homozygous (pure inbred lines), it is possible to cultivate the same genotype on several locations in different years. Thus, by choosing an appropriate field design, the heterogeneity of the residual variance can be

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

greatly reduced. Therefore, in this study, the heterogeneity of the residual variance is not accounted for in the statistical models.

A further increase in prediction accuracy is expected by taking information of molecular markers in the estimation of breeding values into account. One option would be to include genetic similarities calculated based on molecular marker data in the prediction instead of the commonly used genetic relationship matrix (Bauer *et al.*, 2006, 2008). This approach is advantageous if pedigree information among the lines is missing or unavailable. Another strategy is to predict genome-wide breeding values considering the marker scans of the whole genome (Meuwissen *et al.*, 2001).

In self-pollinating populations, the parental lines are characterized by a high degree of inbreeding. In computing the genetic relationship matrix A, Henderson (1976) assumed that the parental base population is not inbred and unrelated. If Henderson's recursive algorithm is also used for parental inbred lines, the resulting breeding values will be underestimated. Hence, in this study the degree of inbreeding of parental lines was taken into account by calculating all diagonal elements of A-matrix by $1+F_i$ (where F_i=coefficient of inbreeding).

A Bayesian approach can be advantageous if the population size is large or the model structure is complex, because it may be easier to compute the sampling of the marginal posterior distributions (Blasco, 2001; Duangjinda *et al.*, 2001) than the frequentist analysis with its two-stage approach. The disadvantage of the Bayesian prediction of breeding value is its high computing costs. Therefore, it is important to sample the marginal posterior distribution efficiently. In general, the Gibbs sampler developed by Geman & Geman (1984) is used. The single-site Gibbs sampler (Gianola & Sorensen, 2002) updates each parameter consecutively, yielding a random sample of the marginal posterior distribution. To obtain a faster convergence rate, García-Cortés & Sorensen (1996) developed a blocked Gibbs sampling algorithm, which we also have used in our study. In this algorithm, the conditional distribution of all parameters is updated in a blockwise manner. However, in a population with large pedigree size the blocked sampler can be slow as solutions to the huge equation systems of the MME are still required. Thus, it could be more efficient to use a hybrid Gibbs sampler, such as that developed by Waldmann *et al.* (2008). In the hybrid sampler, the fast but slow mixing single-site sampler is combined with the slow but fast mixing block updating. Another approach to speed up the blocked Gibbs sampler is to account for rank deficiencies of the coefficient matrix C during matrix inversion. If the C-matrix does not have full rank, a pseudoinverse has to be calculated, which is also computationally demanding. Thus, by considering the nullspace of the C-matrix in the calculation of the pseudoinverse of C, the computing time of sampling the posterior distribution can be shortened, because the nullspace does not change during the MCMC sampling and must therefore be computed only once.

In conclusion, considering the degree of inbreeding in Bayesian analysis was possible without any problems. If genotype-by-environment interactions occur in the data, standard Bayesian Gibbs sampling and a frequentist approach resulted in similar estimates if the trait has a high heritability. For such traits, Bayesian as well as frequentist methods can be recommended although a Bayesian approach could be more appropriate if the statistical model is more complicated. For low heritable traits, however, the standard Gibbs sampling approach is not a suitable strategy to account for genotype-by-environment interactions. Therefore, as related lines are supposed to show similar interactions, an extended genetic relationship matrix was included in the term of genotype-by-environment interaction in the model used to estimate breeding values. This strategy was found to be superior to the commonly used Bayesian model.

Appendix

Assume a symmetric $n\times n$ matrix C having full rank r. In this case, the calculation of C^{-1}, the inverse of C-matrix can be represented by an eigenvalue decomposition of C as

$$C^{-1}=(UDU')^{-1},$$

where D denotes a diagonal matrix with the eigenvalues λ_i as diagonal coefficients and U symbolizes an orthonormal matrix with $U'=U^{-1}$ and $U'U=I$ (with identity matrix I). The columns of U, denoted by $u^{(i)}$, represent the eigenvectors corresponding to the eigenvalues λ_i. Accounting for the orthogonality of the matrix U the inverse matrix C^{-1} is obtained from:

$$C^{-1}=\sum_{i=1}^{n}\frac{1}{\lambda_i}u^{(i)}u^{(i)'}.$$

However, sometimes the number of independent rows or columns of C can be smaller than the total number of rows and columns, which means that a rank deficiency d of the C-matrix has occurred. The rank deficiency can be calculated from $d=n-r$, where r gives the rank of the C-matrix. Assuming $d>0$, the spectral decomposition of C-matrix is as follows:

$$C=(U_1U_2)\begin{pmatrix}D_1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}U_1\\ U_2\end{pmatrix},$$

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press

$$C=(u_1, u_2, \ldots, u_r \quad u_{r+1}, \ldots, u_n)$$

$$\times \left(\begin{array}{ccc:c} \lambda_1 & & & \\ & \cdots & & 0 \\ & & \lambda_r & \\ \hdashline & & & \\ & 0 & & 0 \\ & & & \end{array}\right)$$

$$\times (u_1, u_2, \ldots, u_r \quad u_{r+1}, \ldots, u_n)',$$

$$C=U_1 D_1 U_1' + U_2 0 U_2',$$

$$C=U_1 D_1 U_1',$$

where U_1 and U_2 are orthogonal subspaces of U.

Then, the inverse of the C-matrix can be substituted by the pseudoinverse, which is a special kind of a generalized inverse (Ben-Israel & Greville, 2003). The pseudoinverse of C, denoted as C^+, can be derived from:

$$C^+ = U_1 D_1^{-1} U_1' = \sum_{i=1}^{r} \frac{1}{\lambda_i} u^{(i)} u^{(i)'}.$$

To speed up the computation, the orthogonal nullspace U_2 of the C-matrix can be used in the calculation of the pseudoinverse C^+ (Koch, 2007). For that, the orthogonal subspace U_2 is added to the matrix C resulting in a regular matrix spanning the full space and existing inverse. After the inversion, the subspace U_2 is subtracted. Therefore, the pseudoinverse C^+ can now be represented by

$$C^+ = (C + U_2 U_2')^{-1} - U_2 U_2'$$

or equivalently with

$$C^+ = (U_1 D_1 U_1' + U_2 U_2')^{-1} - U_2 U_2'.$$

Proof.

$$C^+ = (U_1 D_1 U_1' + U_2 U_2')^{-1} - U_2 U_2',$$

$$C^+ = \left(\sum_{i=1}^{r} \lambda_i u^{(i)} u^{(i)'} + \sum_{i=r+1}^{n} u^{(i)} u^{(i)'} \right)^{-1} - \sum_{i=r+1}^{n} u^{(i)} u^{(i)'},$$

$$C^+ = \sum_{i=1}^{r} \frac{1}{\lambda_i} u^{(i)} u^{(i)'} + \sum_{i=r+1}^{n} u^{(i)} u^{(i)'} - \sum_{i=r+1}^{n} u^{(i)} u^{(i)'},$$

$$C^+ = \sum_{i=1}^{r} \frac{1}{\lambda_i} u^{(i)} u^{(i)'} = (U_1 D_1 U_1')^{-1}.$$

In conclusion, using the orthogonal nullspace U_2 of C-matrix in the calculation of the pseudoinverse C^+ speeds up the inversion of C because in our application the subspace U_2 has to be computed only once and does not change during the MCMC sampling.

We gratefully thank Patrik Waldmann (University of Umeå, Sweden) for a Gibbs sampling Matlab code that provided a basis for our developing work and Boby Mathew (University of Bonn, Germany) for helpful discussions about the manuscript. We also thank Shizhong Xu (University of California, USA) and an anonymous reviewer for valuable comments and suggestions on the manuscript.

References

Bauer, A. M. & Léon, J. (2008). Multiple-trait breeding values for parental selection in self-pollinating crops. *Theoretical and Applied Genetics* **116**, 235–242.

Bauer, A. M., Reetz, T. C. & Léon, J. (2006). Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. *Crop Science* **46**, 2685–2691.

Bauer, A. M., Reetz, T. C. & Léon, J. (2008). Predicting breeding values of spring barley accessions by using the singular value decomposition of genetic similarities. *Plant Breeding* **127**, 274–278.

Becker, H. C. & Léon, J. (1988). Stability analysis in plant breeding. *Plant Breeding* **101**, 1–23.

Ben-Israel, A. & Greville, T. N. E. (2003). Generalized Inverses – Theory and Applications. 2nd edn. New York: Canadian Mathematical Society, Springer.

Blasco, A. (2001). The Bayesian controversy in animal breeding. *Journal of Animal Science* **79**, 2023–2046.

Burton, P. R., Tiller, K. J., Gurrin, L. C., Cookson, W. O. C. M., Musk, A. W. & Palmer, L. J. (1999). Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genetic Epidemiology* **17**, 118–140.

Duangjinda, M., Misztal, I., Bertrand, J. K. & Tsuruta, S. (2001). The empirical bias of estimates by restricted maximum likelihood, Bayesian method, and method *R* under selection for additive, maternal, and dominance models. *Journal of Animal Science* **79**, 2991–2996.

Falconer, D. S. & Mackay, T. F. C. (1996). Introduction to Quantitative Genetics. 4th edn. New York: Pearson.

Fan, X., Felsővályi, A., Sivo, S. A. & Keenan, S. C. (2002). SAS for Monte Carlo Studies: A Guide for Quantitative Researchers. Cary, NC: SAS Institute.

Fernando, R. L., Knights, S. A. & Gianola, D. (1984). On a method of estimating the genetic correlation between characters measured in different experimental units. *Theoretical and Applied Genetics* **67**, 175–178.

García-Cortés, L. A. & Sorensen, D. (1996). On a multivariate implementation of the Gibbs sampler. *Genetics Selection Evolution* **28**, 121–126.

Gauch, H. G. (1988). Model selection and validation for yield trials with interaction. *Biometrics* **44**, 705–715.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Gianola, D. & Sorensen, D. (2002). Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. New York: Springer-Verlag.

Gilmour, A. R., Gogel, B. J., Cullis, B. R. & Thompson, R. (2005). ASReml User Guide Release 2.0. Hemel Hempstead, UK: VSN International Ltd.

Gilmour, A. R., Thompson, R. & Cullis, B. R. (1995). AI, an efficient algorithm for REML estimation in linear mixed models. *Biometrics* **51**, 1140–1450.

Gwaze, D. P. & Woolliams, J. A. (2001). Making decisions about the optimal selection environment using Gibbs sampling. *Theoretical and Applied Genetics* **103**, 63–69.

Hanson, W. D. (1963). Heritability. In *Statistical Genetics And Plant Breeding* (ed. W. D. Hanson & H. F. Robinson), pp. 125–139. Washington, DC: National Academy of Sciences – National Research Council.

Harville, D. & Carriquirry, A. (1992). Classical and Bayesian predictions as applied to an unbalanced mixed linear model. *Biometrics* **48**, 987–1003.

Hazelton, M. L. & Gurrin, L. C. (2003). A note on genetic variance components in mixed models. *Genetic Epidemiology* **24**, 297–301.

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**, 69–93.

Henderson, C. R. (1984). Applications of Linear Models in Animal Breeding. Guelph, ON: University of Guelph.

Hoti, F. J., Sillanpää, M. J. & Holmström, L. (2002). A note on estimating the posterior density of a quantitative trait locus from a Markov chain Monte Carlo sample. *Genetic Epidemiology* **22**, 369–376.

Koch, K. R. (2007). Parameter Estimation and Hypothesis Testing in Linear Models. 2nd ed. New York: Springer-Verlag.

Lin, S. (1999). Monte Carlo Bayesian methods for quantitative traits. *Computational Statistics and Data Analysis* **31**, 89–108.

Matlab (2007). High-performance Numeric Computation and Visualization Software, Version 7. Natick, MA: The Math Works Inc.

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

Oakey, H., Verbyla, A., Pitchford, W., Cullis, B. & Kuchel, H. (2006). Joint modelling of additive and non-additive genetic line effects in single field trials. *Theoretical and Applied Genetics* **113**, 809–819.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545–554.

Piepho, H. P., Möhring, J., Melchinger, A. E. & Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* **161**, 209–228.

Quaas, R. L. (1976). Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* **32**, 949–953.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15–51.

SAS Institute (2004). The SAS system for Windows. Release 9.1. Cary, NC: SAS Institute.

Schenkel, F. S., Schaeffer, L. R. & Boettcher, P. J. (2002). Comparison between estimation of breeding values and fixed effects using Bayesian and empirical BLUP estimation under selection on parents and missing pedigree information. *Genetics Selection Evolution* **34**, 41–59.

Smith, A., Cullis, B. & Gilmour, A. (2001). The analysis of crop variety evaluation data in Australia. *Australian and New Zealand Journal of Statistics* **43**, 129–145.

Sorensen, D. A., Wang, C. S., Jensen, J. & Gianola, D. (1994). Bayesian analysis of genetic change due to selection using Gibbs sampling. *Genetics Selection Evolution* **26**, 333–360.

Soria, F., Basurco, F., Toval, G., Silio, L., Rodriguez, M. C. & Toro, M. (1998). An application of Bayesian techniques to the genetic evaluation of growth traits in *Eucalyptus globulus*. *Canadian Journal of Forest Research* **28**, 1286–1294.

Thomas, D. C. (1992). Fitting genetic data using Gibbs sampling – an application to nevus counts in 38 Utah kindreds. *Cytogenetics and Cell Genetics* **59**, 228–230.

Van Tassell, C. P., Casella, G. & Pollak, E. J. (1995). Effects of selection on estimates of variance components using Gibbs sampling and restricted maximum likelihood. *Journal of Dairy Science* **78**, 678–692.

Waldmann, P. & Ericsson, T. (2006). Comparison of REML and Gibbs sampling estimates of multi-trait genetic parameters in Scots pine. *Theoretical and Applied Genetics* **112**, 1441–1451.

Waldmann, P., Hallander, J., Hoti, F. & Sillanpää, M. J. (2008). Efficient Markov Chain Monte Carlo implementation of Bayesian analysis of additive and dominance genetic variances in noninbred pedigrees. *Genetics* **179**, 1101–1112.

Wang, C. S., Rutledge, J. J. & Gianola, D. (1993). Marginal inference about variance components in a mixed linear model using Gibbs sampling. *Genetics Selection Evolution* **21**, 41–62.

Weber, W. E. & Wricke, G. (1990). Genotype × environment interaction and its implications in plant breeding. In *Genotype-by-environment Interaction and Plant Breeding* (ed. M. S. Kang), pp. 1–19. Baton Rouge, LA: Louisiana State University.

Zeger, S. & Karim, M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **26**, 79–86.

Zeng, W., Ghosh, S. & Li, B. (2004). A blocking Gibbs sampling method to detect major genes with phenotypic data from a diallel mating. *Genetical Research* **83**, 143–154.

https://doi.org/10.1017/S0016672309000160 Published online by Cambridge University Press