Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 1037–1046 doi: 10.1111/2041-210X.12113

Rapid Bayesian inference of heritability in animal models without convergence problems

Jon Ahlinder[1] **and Mikko J. Sillanpää**[2,3,4*]

[1]Division of CBRN Defence and Security, Swedish Defence Research Agency (FOI), Umeå SE-901 82,Sweden; [2]Department of Mathematical Sciences, Department of Biology and Biocenter Oulu, University of Oulu, Oulu FIN-90014, Finland; [3]Department of Mathematics and Statistics, University of Helsinki, Helsinki FIN-00014, Finland; and [4]Department of Agricultural Sciences, University of Helsinki, Helsinki FIN-00014, Finland

Summary

1. The rapid advancement in genome sequencing techniques allows the dissection of complex traits of non-model organisms of importance in evolutionary biology, conservation genetics, breeding and medicine. This advancement requires new statistical analysis tools that can handle large amount of sequencing data efficiently.

2. We propose an analytic Bayesian implementation of the mixed linear model which allows rapid and robust inferences of heritability. The two main features of the method are (i) breeding values and residual variance component are analytically integrated out of the model and (ii) the parameter space of the variance ratio parameter is discretized so that a Gibbs sampling distribution can be utilized. We propose further two separate methods to infer breeding values that acknowledge uncertainty of the learned heritability. The benefit of the method compared to a standard Markov Chain Monte Carlo (MCMC) method is visualized on public data sets: two simulated data sets and one Wheat (*Triticum aestivum* L.) pedigree.

3. Results show that while the accuracy of inferred heritability obtained by the proposed and standard methods are almost identical, the computational performance is very encouraging: up to hundred fold speed up and the possibility to make parallel implementation is particularly appealing here, which may further speed up computations.

4. The method allows analysis using a non-invertible relationship matrix so that *ad hoc* manipulation is avoided which can be important as our results imply. We completely avoid convergence and mixing problems here: this is a well-known problem of MCMC simulation, which sometimes can severely reduce the inferential power. Bayes factors for model comparisons can be conveniently calculated as a by-product of the inference procedure. The source code will be available for download at http://www.rni.helsinki.fi/~mjs.

Key-words: analytic Bayesian inference, Bayes factors, breeding value, complex trait analysis, Gibbs sampling distribution, Markov Chain Monte Carlo, mixed-effects models, SNP, *Triticum aestivum*

Introduction

One fundamental population parameter of interest in ecology, in medical genetics, in breeding and conservation genetics, and in evolutionary biology is heritability (Lynch & Walsh 1998; Visscher, Hill & Wray 2008). In a population under study, if heritable genetic variation underlies the trait of interest, a response to natural or artificial selection is expected. This will alter the distribution of phenotypes in the population so that changes due to selection are passed on to future generations. Changes in selection pressure could, for example, involve ecological factors such as climate changes which, in turn, can have evolutionary implications.

Statistical methods for estimating heritability (and breeding values, BV) have therefore received much attention in the quantitative genetic literature (Meuwissen, Hayes & Goddard 2001; Sorensen & Gianola 2002; Thompson 2008; Sillanpää 2011). One example of a convenient and popular method is the animal model, which has been utilized during many decades in the field of animal breeding (Henderson 1975, 1984; Wang, Rutledge & Gianola 1993). The animal model (i.e. a mixed linear model) combines individual phenotypic records with pedigree and/or genetic marker information to infer parameters of interest. Typically, the pedigree information is incorporated into the form of the additive genetic relationship matrix, $\mathbf{A}_P$ (subscript P stands for pedigree), which is included as a covariance matrix in the mixed model analysis. Either, pedigrees are known as in controlled breeding designs or inferred indirectly based on genetic marker data through relatedness estimators

[*]Correspondence author. E-mail: mjs@rolf.helsinki.fi

(Ritland 2000; Pemberton 2008; Riester, Stadler & Klemm 2009). Recently, there has been a rise in the use of animal models to analyse data collected from wild populations (Kruuk 2004; Brommer, Rattiste & Wilson 2008; Frentiu *et al.* 2008). Another area of application is genomewide association studies where the animal model has been used to correct false-positive association due to cryptic relatedness in the analysed population (Yu *et al.* 2006; Aulchenko, de Koning & Haley 2007; Kang *et al.* 2010) and to infer the 'missing' heritability from associated SNPs (Yang *et al.* 2010; Golan & Rosset 2011; Sillanpää 2011).

The Bayesian paradigm has recently gained popularity in complex trait studies (Gianola & van Kaam 2008; Gasbarra *et al.* 2009; Crossa *et al.* 2010; Hallander *et al.* 2010; Steinsland & Jensen 2010; Mathew *et al.* 2012). An appealing property of Bayesian methods is that parameter uncertainty is naturally incorporated in the analysis. Since a probabilistic framework is adopted, output is given as probability distributions which are easy to interpret and credible regions can directly be obtained without the need for making asymptotic assumptions. To obtain estimated marginal posterior distributions of the unknown parameters in the statistical model, a class of powerful methods named Markov Chain Monte Carlo (MCMC) have been successfully employed (Gilks, Spiegelhalter & Richardson 1995). To draw inferences in animal models, MCMC methods have been widely used since the early/mid-1990s, for example, the standard additive polygenic model (Wang, Rutledge & Gianola 1993), the non-additive genetic model (Waldmann *et al.* 2008; Mathew *et al.* 2012), interaction models such as genotype by environment (Bauer *et al.* 2009). One particular advantage of Bayesian inference methods in animal model frameworks is that scale and location parameters are jointly inferred and uncertainty is thus acknowledged, as opposed to the frequentist counterpart, where maximum likelihood estimates of variance components are first obtained and then used as though the point estimates were the true values, to obtain best linear unbiased prediction (BLUP) of BV (Sorensen & Gianola 2002). In addition, there is no need to find good starting values as required in restricted maximum likelihood (REML) techniques which might have an impact on the convergence of the algorithm (Piepho *et al.* 2012).

Recent breakthroughs in molecular genetics have made dense marker panels available in many non-model species of interest to ecologists (Santure *et al.* 2010), breeders (Resende *et al.* 2012) and human geneticists (International HapMap Consortium 2007). These marker panels can be utilized to infer genomic relationships in the animal model framework by estimating the realized or genomic relationship matrix, $\mathbf{A}_G$ (denoted $\mathbf{G}$ in VanRaden 2008), to be used in place of $\mathbf{A}_P$, as for example shown by VanRaden (2008), Daetwyler *et al.* (2010) and Resende *et al.* (2012). The elements in $\mathbf{A}_G$ contain the realized proportion of the genome that is identical by descent (IBD) between pairs of pedigree members. Estimating this proportion of IBD requires sufficient marker coverage of genotyped individuals. Even though the idea of using dense marker panels in animal models is very promising, some hurdles remain to be overcome.

First, the required computational time is unfortunately massive. Most traditional methods for drawing inferences in animal models rely on sparse solvers, since most entries in the pedigree-derived $\mathbf{A}_P^{-1}$ are zero (Henderson 1984). With the introduction of dense marker panels, however, most pairwise relationships in $\mathbf{A}_G^{-1}$ will be nonzero, making sparse matrix techniques unpractical and slow from a computational perspective, as computing the likelihood requires substantial efforts (Legarra & Misztal 2008). Truncating relationships close to zero in $\mathbf{A}_G^{-1}$ would increase sparseness but at the expense of introducing biases (i.e. an ill-defined, non-convex likelihood surface) and numerical instability which could cause convergence problems. When applying MCMC methods to draw inferences, in particular, the standard single-site Gibbs sampler (Sorensen & Gianola 2002), the high posterior correlation may in some case cause mixing problems and thus prevent converge of the MCMC. This problem may require a large number of iterations which typically is time-consuming. In addition, estimation of posterior distribution based on MCMC sampling is done from dependent samples which may have reduced accuracy (due to low effective sample size) when there is lot of dependence among the samples (i.e. for non-sparse data).

Secondly, standard pairwise relationship estimation methods may cause $\mathbf{A}_G$ to be singular and therefore non-invertible. This makes mixed model analysis problematic and *ad hoc* methods might be needed to make the matrix invertible which, in turn, might lead to biased genetic parameter estimates. VanRaden (2008) suggested that a small proportion of $\mathbf{A}_P$, which is always invertible for known pedigrees, could be added to $\mathbf{A}_G$ to avoid the singularity problem. One obvious drawback, apart from the possible introduction of bias, is that pedigrees are seldom known in wild populations and $\mathbf{A}_P$ is, therefore, not available. Alternative strategies to avoid singularity have been proposed, such as ridge regression or G-BLUP (Piepho 2009), variable transformations (Piepho *et al.* 2012), matrix bending techniques (Maenhout, DeBaets & Haensert 2009) and reducing the rank by spectral decomposition (Frentiu *et al.* 2008). Although these approaches have shown to improve numerical stability, they result in approximate inferences (matrix bending and reduced rank decomposition) or depend on user input (i.e. fine tuning of input in *ad hoc* manipulation). Piepho *et al.*'s (2012) suggestion of a transformation of the random genetic effects, which makes it possible to infer the heritability without inverting the original relationship matrix (Waldmann *et al.* 2008), results in exact inference, as well as the ridge regression technique. The problem of developing models to infer heritability for non-definite relationship matrices in a computationally efficient way requires more attention.

The aim of the current paper is to develop a rapid method for analysing large marker data of quantitative traits and make inferences of BV and heritability. The presented method consists of two main steps: first, location parameters and the residual variance component are analytically integrated out of the likelihood. The range of values for the remaining parameter (a ratio of genetic and residual variance) is discretized so that a discrete fully conditional (Gibbs sampling) distribution can be rapidly calculated for obtaining posterior probabilities at different values of the variance ratio, which is proportional to the heritability of the pedigree. BV can then be obtained as a second step by standard sampling-based MCMC procedures and uncertainty in inferred heritability is taken into account. In order to visualize the improvement in speed of the developed model, two simulated pedigrees and a real Wheat pedigree, previously published by Lund *et al*. (2009), Meuwissen & Goddard (2010; shown in Appendix S1) and Crossa *et al*. (2010) with dense marker maps available, are analysed and results are compared with those obtained from a traditional MCMC method (Sorensen & Gianola 2002) and a REML method (Meyer 2007). We show how a model selection analysis can be executed in order to evaluate competing genetic relationship structures (shown in Appendix S2). Additional sensitivity analyses are shown in Appendix S3. Finally, the two random effects case is shown in Appendix S4 for joint inference of heritability and dominance genetic proportion.

Materials and methods

STATISTICAL MODEL

With Gaussian assumptions, we made use of the following linear mixed effect model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \qquad \text{eqn 1}$$

where $\mathbf{y}$ is a vector of size $n \times 1$ containing phenotypic records of a continuous trait for all members in the population. Following the Bayesian view (Sorensen & Gianola 2002, pp. 313), fixed effects are treated as random and are considered to have distributional assumptions. Thus, $\mathbf{b}$ is a vector of size $p \times 1$ containing systematic environmental effects (i.e. fixed effects) that follows a multivariate normal distribution with zero mean vector, and prior covariance matrix $\mathbf{B}\sigma_b^2$, where $\mathbf{B}$ is a non-singular unscaled covariance matrix of size $p \times p$ and σ_b^2 is the scale parameter. Here, $\mathbf{B}\sigma_b^2$ is treated as known. $\mathbf{u}$ is a vector of size $n \times 1$ containing genetic effects that follow a multivariate normal distribution with zero mean vector, and covariance structure $\mathbf{A}_G\sigma_u^2$, where $\mathbf{A}_G$ is the genomic relationship matrix of size $n \times n$ and σ_u^2 is the genetic variance component. Known incidence matrices $\mathbf{X}$ and $\mathbf{Z}$ are relating phenotypic records to respective location parameters included in (1), and $\mathbf{e}$ is a vector containing independent residual errors that follow a multivariate normal distribution with zero mean vector, and covariance structure $\mathbf{I}\sigma_e^2$, where $\mathbf{I}$ is the identity matrix of order n. Throughout the paper, we will use the one genetic (random) effect case in all equations, but it is straight forward to generalize the model to handle multiple random effects. See Appendix S4 for how two random effects could be handled to rapidly infer heritability and dominance genetic proportion for the real Wheat pedigree.

In the present paper, we propose a two-step approach for rapid inference of the parameters in the animal model (1). If $\mathbf{y}$ is assumed to follow a Gaussian distribution, with $\mathbf{e}$ identically and independently distributed, according to Sorensen & Gianola (2002), the resulting likelihood function is

$$p(\mathbf{y}|\mathbf{b},\mathbf{u},\sigma_e^2) = (2\pi)^{-n/2}\sigma_e^{-1} \exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{u})^T(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{u})\right\}. \qquad \text{eqn 2}$$

The joint posterior density of all unknown parameters is proportional to the likelihood multiplied with the prior distribution of the unknown parameters in the hierarchical model according to

$$p(\mathbf{b},\mathbf{u},\sigma_b^2,\sigma_u^2,\sigma_e^2|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{b},\mathbf{u},\sigma_e^2)p(\mathbf{b}|\sigma_b^2)p(\mathbf{u}|\sigma_u^2)p(\sigma_b^2)p(\sigma_u^2)p(\sigma_e^2) \qquad \text{eqn 3}$$

In order to derive a more parsimonious model, one trick is to perform marginalization of (3), where both $\mathbf{b}$ and $\mathbf{u}$ can be treated as nuisance parameters and, consequently, be integrated out from the hierarchical model (Searle, Casella & McCulloch 1992). The following marginal density can then be obtained

$$p(\sigma_b^2,\sigma_u^2,\sigma_e^2|\mathbf{y}) = \int p(\mathbf{b},\mathbf{u},\sigma_b^2,\sigma_u^2,\sigma_e^2|\mathbf{y})\mathrm{d}\mathbf{b}\mathrm{d}\mathbf{u} \propto p(\sigma_b^2)p(\sigma_u^2)p(\sigma_e^2) \int p(\mathbf{y}|\mathbf{b},\mathbf{u},\sigma_e^2)p(\mathbf{b}|\sigma_b^2)p(\mathbf{u}|\sigma_u^2)d\mathbf{b}d\mathbf{u}. \qquad \text{eqn 4}$$

Here, proper prior distributions for $\mathbf{b}$ and $\mathbf{u}$ need to be specified, such as $\mathbf{b}|\mathbf{B},\sigma_b^2 \sim \mathrm{MVN}(\mathbf{0},\mathbf{B}\sigma_b^2)$ and $\mathbf{u}|\mathbf{A}_G,\sigma_u^2 \sim \mathrm{MVN}\ (\mathbf{0},\mathbf{A}_G\sigma_u^2)$: integration over these distributions are one. The integration over the likelihood function results in

$$\begin{aligned} p(\mathbf{y}|\Sigma) &= \int p(\mathbf{y}|\mathbf{b},\mathbf{u},\sigma_e^2)p(\mathbf{b}|\sigma_b^2)p(\mathbf{u}|\sigma_u^2)d\mathbf{b}d\mathbf{u} \\ &= (2\pi)^{-n/2}\det(\Sigma)^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y}\right\}, \end{aligned} \qquad \text{eqn 5}$$

where $\Sigma = \mathbf{X}\mathbf{B}\mathbf{X}^T\sigma_b^2 + \mathbf{Z}\mathbf{A}_G\mathbf{Z}^T\sigma_u^2 + \mathbf{I}\sigma_e^2$. This likelihood function (5) does not contain $\mathbf{b}$ and $\mathbf{u}$ and is similar to that presented by several authors (Sorensen & Gianola 2002, pp. 313–316; Aulchenko, de Koning & Haley 2007). In the corresponding REML likelihood, BV are marginalized and fixed effects are made orthogonal (i.e. having no influence), which is an analogous operation (Thompson 2008). The obtained joint posterior distribution can be written as

$$p(\sigma_b^2,\sigma_u^2,\sigma_e^2|\mathbf{y}) \propto p(\mathbf{y}|\sigma_b^2,\sigma_u^2,\sigma_e^2)p(\sigma_b^2)p(\sigma_u^2)p(\sigma_e^2). \qquad \text{eqn 6}$$

Here, we assume that the analyst will pre-specify the prior variance of systematic environmental effects so that $p(\sigma_b^2) = 1$.

ANALYTIC INTEGRATION OF RESIDUAL VARIANCE COMPONENT FROM THE LIKELIHOOD

Gasbarra *et al*. (2009) showed how to integrate σ_e^2 out of the likelihood of a model of the form $\mathbf{y} \sim \mathrm{MVN}(\mathbf{0},\Sigma)$ (O'Hagan & Forster 2004, Ch. 11). We assign an inverse gamma prior to σ_e^2, which is a convenient choice, since this prior is the conjugate prior distribution for the normal variance (Gelman *et al*. 2004). In the present paper, the covariance matrix, Σ, can be rewritten as

$$\Sigma = \sigma_e^2(\mathbf{X}\mathbf{B}\mathbf{X}^T\frac{\sigma_b^2}{\sigma_e^2} + \mathbf{Z}\mathbf{A}_G\mathbf{Z}^T\frac{\sigma_u^2}{\sigma_e^2} + \mathbf{I}) = \sigma_e^2\Sigma^\star. \qquad \text{eqn 7}$$

For simplification, let $\lambda_u = (\sigma_e^2)/(\sigma_u^2)$ and $\lambda_b = (\sigma_e^2)/(\sigma_b^2)$ so that $\Sigma = \sigma_e^2\Sigma^\star = \sigma_e^2(\mathbf{X}\mathbf{B}\mathbf{X}^T\lambda_b^{-1} + \mathbf{Z}\mathbf{A}_G\mathbf{Z}^T\lambda_u^{-1} + \mathbf{I})$. The heritability h^2 can be expressed as a function of λ_u by combining $h^2 = (\sigma_u^2)/(\sigma_p^2)$, $\lambda_u = (\sigma_e^2)/(\sigma_u^2)$ and $\sigma_p^2 = \sigma_u^2 + \sigma_e^2$, which gives

$$h^2 = \frac{1}{(1+\lambda_u)}. \qquad \text{eqn 8}$$

Following Gasbarra *et al*. (2009), the likelihood is obtained as

$$p(\mathbf{y}|\Sigma^\star) = (2\pi)^{-n/2}\det(\Sigma^\star)^{-1/2}\frac{(a/2)^{d/2}\Gamma((d+n)/2)}{(a^\star/2)^{(d+n)/2}\Gamma(d/2)}, \qquad \text{eqn 9}$$

where $a^{\star} = a + \mathbf{y}^T(\Sigma^{\star})^{-1}\mathbf{y}$, a and d are hyperparameters of the inverse gamma prior for σ_e^2 and $\Gamma(.)$ is the gamma function. We assume that λ_b^{-1} is a constant having a large value (i.e. $\lambda_b^{-1} = 1000$), which reflects uninformative prior knowledge of group level effects as $\sigma_b^2 \gg \sigma_e^2$. Throughout the analysis, the λ_b^{-1} is kept constant: this guarantees a uniform prior which resembles what is assumed for fixed effects in a classic REML analysis. In making this assumption, we assume in addition that λ_b^{-1} and λ_u^{-1} are mutually independent. A similar assumption was made by Gasbarra *et al.* (2009) for QTL and polygenic variance ratios. Hence, the only unknown parameter left in our model is λ_u, and the corresponding joint posterior distribution is

$$p(\lambda_u^{-1}|\mathbf{y}) \propto p(\mathbf{y}|\lambda_u^{-1})p(\lambda_u^{-1}). \qquad \text{eqn 10}$$

CALCULATING DISCRETE GIBBS SAMPLING DISTRIBUTION FOR LAMBDA

In order to speed up the genetic analysis, parameter space of λ_u^{-1} is discretized on a finite number of categories in the range of interest. As we only have one unknown parameter in the model (λ_u^{-1}), the Gibbs sampling distribution equals directly the posterior distribution. Bayesian inference of λ_u is given by Bayes theorem, where the posterior probability of the *j*th single category can be written as

$$p(\lambda_{u,j}^{-1}|\mathbf{y}) = \frac{p(\mathbf{y}|\lambda_{u,j}^{-1})p(\lambda_{u,j}^{-1})}{\sum_{k=1}^{N} p(\mathbf{y}|\lambda_{u,k}^{-1})p(\lambda_{u,k}^{-1})}, \qquad \text{eqn 11}$$

where N is the total number of categories in the range of λ_u^{-1} (i.e. $RANGE(\lambda_u^{-1}) = [p(\lambda_{u,1}^{-1}|\mathbf{y}), p(\lambda_{u,2}^{-1}|\mathbf{y}), \ldots, p(\lambda_{u,N}^{-1}|\mathbf{y})]$). All N probabilities are computed so that a discrete posterior distribution is obtained for λ_u. The denominator of (11), the marginal likelihood, is the normalizing constant, which is an important part in Bayesian model selection (Kass & Raftery 1995). In Appendix S2, we show a simple approach to compare marginal likelihoods of competing models.

We used the following prior probability: $p(\lambda_u^{-1}) \sim U(0,4)$, which corresponds to a range of h^2 covering most applications in quantitative trait analysis ($0 \leq h^2 \leq 0{\cdot}8$). Note that if $h^2 \to 1$, $\lambda_u^{-1} \to \infty$. In Appendices S3 and S4, we show an alternative discretization of h^2 directly, which allows the entire parameter space to be evaluated (i.e. $0 \leq h^2 \leq 1{\cdot}0$). Conditional posterior probabilities are obtained by combining the likelihood (9) and prior U(0,4) using (11). In order to calculate $\det(\Sigma^{\star})$, standard formulas were used (Golub & van Loan 1996).

INFERENCES OF POSTERIOR DISTRIBUTION FOR LOCATION PARAMETERS

The posterior distribution of location parameters, $\theta = [\mathbf{b}, \mathbf{u}]^T$, is a mixture distribution according to

$$\begin{aligned} p(\mathbf{b},\mathbf{u}|\lambda_b^{-1},\mathbf{y}) = {} & p(\mathbf{b},\mathbf{u}|\lambda_b^{-1},\lambda_u^{-1},1 = l_u,\mathbf{y})p(\lambda_u^{-1},1 = l_u|\mathbf{y}) \\ & + p(\mathbf{b},\mathbf{u}|\lambda_b^{-1},\lambda_u^{-1},2 = 2l_u,\mathbf{y})p(\lambda_u^{-1},2 = 2l_u|\mathbf{y}) + \ldots + \\ & + p(\mathbf{b},\mathbf{u}|\lambda_b^{-1},\lambda_u^{-1},N = Nl_u,\mathbf{y})p(\lambda_u^{-1},N = Nl_u|\mathbf{y}), \end{aligned} \qquad \text{eqn 12}$$

where l_u is the bin size for λ_u^{-1} (equal bin size is assumed here, but see Appendix S3 for a model with assigned prior on h^2, which results in unequal bin size for λ_u^{-1}), $p(\lambda_b^{-1}|\mathbf{y}) = 1$ and is omitted from (12). Note that $p(\mathbf{b},\mathbf{u}|\lambda_b^{-1},\mathbf{y})$ is marginalized over λ_u^{-1} in (12). The posterior mean of the above mixture distribution (12) can be obtained by the following set of equations (i.e. the Bayesian version of Henderson's mixed model equations)

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} + \mathbf{B}^{-1}\lambda_b & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \mathbf{A}_G^{-1}\lambda_{u,i} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}, \qquad \text{eqn 13}$$

where index i refers to the *i*th bin in the discrete lambda distribution. The coefficient matrix on the left hand side in (13) is denoted $\mathbf{C}$. In addition, the solution of linear system (13) provides the mean of the fully conditional posterior distribution of the location parameters (Sorensen & Gianola 2002). As λ_b^{-1} is set to an arbitrary large constant (vague knowledge of group level effects), we only need $\lambda_{u,i}, i = 1\ldots N$, in order to obtain posterior mean of $\mathbf{b}$ and $\mathbf{u}$. Thus, to obtain conditional expectations (CE) of $\mathbf{b}$ and $\mathbf{u}$, the approach is fully Bayesian and no approximations is introduced. To obtain the predicted error variance (PEV) of each location parameter, we need to introduce approximations into our approach, since the posterior distribution of the error variance component, σ_e^2, is needed (i.e. needs to be separated from σ_u^2 in λ_u^{-1}). First, we need an estimate of the group level effect, $\hat{\mathbf{b}}$, which can be obtained using ordinary least square (OLS) technique: $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ (Lynch & Walsh 1998, p. 200). Then, we estimate the empirical phenotypic variance using $\hat{\sigma}_y^2 = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T\mathbf{Z}\mathbf{A}_G\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})/(n-1)$ and compute the posteriors of genetic and residual variance as $\sigma_u^2 = h^2\hat{\sigma}_y^2$ and $\sigma_e^2 = \hat{\sigma}_y^2 - \sigma_u^2$. As this step does not involve the posterior distribution of σ_y^2 and $\mathbf{b}$, but point estimates, this is an empirical Bayes estimation step. Furthermore, we acknowledge that OLS estimates can be sensitive to small pedigree sizes, unbalanced mating designs in artificial populations and presence of selection bias. PEV is calculated by extracting the diagonal of the inverted coefficient matrix, $\mathbf{C}^{-1}$, in (13) and multiply it with σ_e^2. We will denote this method, to infer posterior distribution for location parameters, CE.

If the full posterior of BV is of inferential interest, an MCMC Gibbs sampler could be applied. By estimating the posteriors of σ_u^2 and σ_e^2, as mentioned above, we would utilize (12) as a mixture distribution for obtaining the estimated posterior $p(\theta|\mathbf{y})$ after replacing λ_u^{-1} and λ_b^{-1} with the corresponding scale parameters. The only parameters to be updated in the model are the location parameters, as posteriors of the scale parameters are already estimated (through λ_u^{-1}, where σ_u^2 and σ_e^2 needs to be separated by an OLS estimate of σ_y^2 as above). The conditional posterior distribution of the location parameters is obtained from Sorensen & Gianola (2002) as $\theta|\sigma_u^2, \sigma_e^2, \mathbf{y} \sim \text{MVN}(\hat{\theta}, \mathbf{C}^{-1}\sigma_e^2)$. Since samples are drawn from the discrete posterior (Gibbs) distribution which makes parameters independent, we do not need any burn-in and only a small to moderate number of MCMC iterations is needed: the size of the chain is proportional to the inverse of the standard errors caused by the Monte Carlo procedure. The number of MCMC iterations was 5000 throughout the study. We denote this approach blMCMC, which stands for blocked MCMC. It should be pointed out that this approach is not fully Bayesian as it involves the point estimates of σ_y^2 and $\mathbf{b}$. Note that we re-estimate $\mathbf{b}$ by solving (13).

REFERENCE PARAMETER ESTIMATION METHODS

We implemented both the standard single-site Gibbs and the blocked Gibbs sampler, as shown in Sorensen & Gianola (2002). In single-site sampling, each parameter is drawn from its fully conditional posterior distribution, $[\theta_i|\theta_{-i}, \sigma_u^2, \sigma_e^2, \mathbf{y}]$, where θ_{-i} is a vector containing all location parameters except θ_i. The two variance components σ_u^2 and σ_e^2 are drawn from scaled inverted chi-squared distributions and are assumed to be conditionally independent of the location parameters $\theta = [\mathbf{b}^T, \mathbf{u}^T]^T$. The implemented single-site algorithm is described in Sorensen & Gianola (2002, pp. 566–570). Here, the chains were run for 225 000 iterations with a thinning of 10 and a burn-in of 25 000 leaving the final Markov chain to 20 000 samples. In blocked Gibbs sampling, θ is jointly drawn in a blockedwise way, $\theta|\sigma_u^2, \sigma_e^2, \mathbf{y}$ given previously sampled values of σ_u^2 and σ_e^2, which are separately drawn (Garcia-Cortes & Sorensen 1996). Our implementation differ from the algorithm suggested by Garcia-Cortes & Sorensen (1996; Sorensen & Gianola 2002, pp. 587–588) in that $\hat{\theta}$ are obtained by a direct method using Cholesky decomposition instead of an iterative method, where inversion of **C** is avoided. For example, in Waldmann *et al.* (2008), a Conjugate gradient iterative method was implemented to obtain $\hat{\theta}$ in the blocked sampling step, which is likely to be faster than the current implementation. For blocked sampling, the chain was run for 25 000 iterations with a thinning of 2 and a burn-in of 5000 samples which results in a chain of 10 000 samples. The heritability was computed as a function of the MCMC for the variance components as $h^2 = (\sigma_u^2)/(\sigma_u^2 + \sigma_e^2)$. The standard MCMC was used for timing comparisons as the Bayesian methods were implemented using the same numerical routines for solving equation systems (CLAPACK) and written in the same programming language (ANSI C).

Furthermore, in order to verify the results obtained with our implemented Bayesian methods, we made use of the publicly available software package WOMBAT (Meyer 2007). WOMBAT fits linear mixed effect models through REML. In all comparisons between the REML and Bayesian methods, we used identical group level factors and relationship structures in the animal model.

ANALYSED SYNTHETIC DATA 1

In the present study, we have analysed two simulated pedigrees, typical for animal breeding stocks, where the first data were published by Lund *et al.* (2009) and are freely available on http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html. In total, 5865 pedigree members from seven generations were simulated, where both pedigree and phenotype information are available of individuals for the first to the fourth generation and SNP data are available for all individuals. The trait was controlled by 48 QTLs, and 6000 SNPs were covering six chromosomes at a distance of 0·1 cM between markers (i.e. 1000 SNPs per chromosome). The simulated heritability of the pedigree was $h^2 = 0{\cdot}3$. The genomic relationship matrix was computed using the second method proposed by VanRaden (2008). The observed allele frequencies (p_i) from the first generation in the current population (i.e. the first 165 pedigree members as ordered in the pedigree file) were used in the calculations. SNP genotypes are coded as 1, 0 and −1 for the first arbitrary homozygote (i.e. allele value 2 in the data file), heterozygote and second arbritrary homozygote, respectively. Because the resulting $\mathbf{A}_G$ was not positive definite and, hence, non-invertible, we added a small fraction of the pedigree-derived relationship matrix, $\mathbf{A}_P$, so that $\mathbf{A}_G^\star = 0{\cdot}99\mathbf{A}_G + 0{\cdot}01\mathbf{A}_P$ (VanRaden 2008). As fixed effect, the sex of each member was used.

ANALYSED WHEAT DATA 2

The real data set is a collection of 599 historical CIMMYT Wheat (*Triticum aestivum* L.) lines included in the global wheat breeding programme and previously published by Crossa *et al.* (2010). The phenotype analysed here was the 2-year average grain yield of each of these lines, standardized to a unit variance. For simplicity, we averaged the phenotypes over four different environments. In total, 1279 Diversity Array Technology (DArT) markers were available in the analysis after removing markers with minor allele frequency < 0·05 (i.e. 1447 markers prior to exclusion). These markers are binary, denoted by their presence (1) or absence (0) in the genome. In addition, the pedigree of the breeding population was available so that the additive relationship matrix **A** among the 599 lines could be computed (see http://cropwiki.irri.org/icis/index.php/TDM_GMS_Browse). The realized relationship matrix $\mathbf{A}_G$ was calculated based on the DArT markers using the same method as in the aforementioned examples. However, in order to make $\mathbf{A}_G$ invertible, we added a fraction of $\mathbf{A}_P$: $\mathbf{A}_G^\star = 0{\cdot}99\mathbf{A}_G + 0{\cdot}01\mathbf{A}_P$, as when analysing the example data 1. The statistical model used was the same as in the analysis of data set 2: $\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \mathbf{e}$, although the dimension of vectors **y**, **1**, **u**, **e** was 599 × 1. Crossa *et al.* (2010) obtained a average point estimate of $h^2 = 0{\cdot}353$ averaged over the four environments.

Results

ANALYSED SYNTHETIC DATA 1

Table 1 shows summary statistics obtained from the analysis of pedigree 1 using both the analytic and standard MCMC Gibbs sampling approaches. Here, we report h^2 directly and not λ_u^{-1}, as point estimates are straight forward to calculate using (8). Our posterior point estimates of h^2 and their 95% credible interval (CI) regions closely agreed with those obtained by the standard Gibbs samplers for 100, 250 and 1000 bins. This finding is strengthened by the low level of Kullback–Leibler (K–L) divergence (Kullback & Leibler 1951) of the inferred posteriors and almost equal correlations of mean posteriors of BV to true breeding values (TBV), as reported in Table 1 and seen in Fig. 1. Note that all sets of bins gave almost identical results, in particular mean and the standard deviation of inferred posterior of h^2 and correlations with both obtained by standard Gibbs samplers and the TBV. Hence, in the current example, there is no need to use more than 100 bins to span the parameter space of h^2. In addition, to obtain inferred BV, both suggested approaches (i.e. CE) and blocked MCMC (blMCMC), resulted in equal correlations to both TBV and posterior mean of BV obtained by single-site MCMC. The estimated h^2 obtained by the REML method and results reported by Strandén & Christensen (2011) agreed closely to the mean of inferred posterior obtained by the Bayesian approaches. The blocked Gibbs sampling method was very computationally intense but resulted in similar point estimates as the single-site method.

The computational time required for both approaches varied greatly depending on the number of bins in the analytic approach. For 100 bins, the analytic approach outperformed the single-site MCMC by a factor of 25- to 4-fold reduced computational time at the same accuracy. For 1000 bins, however, the required computational time was marginally better for the analytic approach. The inference method to obtain estimates of BV resulted in similar computational time, although CE seemed more beneficial for a fewer number of bins, whereas blMCMC seemed favoured by a larger number of bins. Reducing the number of MCMC iterations in blMCMC from 5000 to 1000 slightly reduced the correlation with TBV: for 100 bins, $cor(\mathrm{BV}_{1000}, \mathrm{TBV}) = 0{\cdot}863$ compared to $cor(\mathrm{BV}_{5000}, \mathrm{TBV}) = 0{\cdot}865$ (Table 1). On the other hand, the computational time was much reduced: $t_{1000} = 47{\cdot}30$ min compared to $t_{5000} = 136{\cdot}33$ min. All analyses were carried out on an Intel(R) Core(TM)2 Duo CPU processor (2·26 GHz) with

Table 1. Analysed data 1

		Heritability					Correlations		Computational time				
Model	nbins	Mode	Mean	SD	95% CI	K–L	*cor* (BV, TBV)	*cor* (BV$_A$, BV$_{SM}$)	Δt_C	Δt_M	t_{h^2}	$t_{BV,C}$	$t_{BV,M}$
Analytic	100	0·351	0·351	0·022	[0·296, 0·383]	0·013	0·865	1·000	155·83	207·05	70·62	85·21	136·43
Analytic	250	0·347	0·351	0·022	[0·304, 0·391]	0·014	0·865	1·000	359·76	297·13	155·79	203·97	141·34
Analytic	1000	0·348	0·351	0·022	[0·307, 0·394]	0·020	0·865	1·000	1346·80	773·80	552·10	794·70	221·70
MCMC1	–	0·351	0·347	0·023	[0·304, 0·391]	–	0·863	–	–	1630·76	–	–	–
MCMC2	–	0·350	0·348	0·023	[0·302, 0·394]	–	0·863	–	–	~ 37 days	–	–	–
REML	–	0·348	–	–	–	–	0·865	–	–	–	–	–	–

Summary statistics of inferred h^2 obtained from analysis of data set 1, MCMC1 and MCMC2 are the single-site and blocked Gibbs sampling methods, respectively, nbins is the number of bins of h^2, SD is the standard deviation, CI is the credible interval, K–L is the Kullback–Leibler divergence of inferred posteriors of h^2 obtained from analytic and standard MCMC methods. Correlation between inferred breeding values (BV) using the various Bayesian and restricted maximum likelihood (REML) approaches, and true breeding values (TBV) is denoted as *cor*(BV, TBV). Correlation between inferred BV using the analytic approach and standard single-site Gibbs sampler is denoted *cor*(BV$_A$, BV$_{SM}$) for both conditional expectations (CE) and blMCMC inference methods, as both methods resulted in equal correlations. The total computational time for the analytic approach with either CE or blMCMC is denoted Δt_C and Δt_M, respectively. The computational time for the heritability estimation, CE and blMCMC to obtain inferred BVs is denoted t_{h^2}, $t_{BV,C}$ and $t_{BV,M}$. All time units are given in minutes. The parameter range of h^2 is between 0 and 0·8.

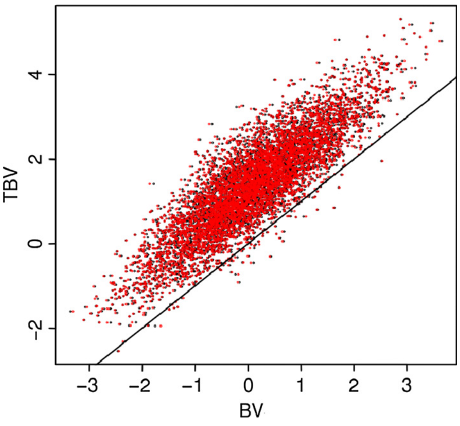

Fig. 1. Correlations of mean of inferred breeding values (BV) and true breeding values (TBV) in data set 1 obtained with analytic and standard Markov Chain Monte Carlo approaches. The black circles are correlations obtained by the analytic approach using conditional expectations to infer BV. The red circles are obtained by the single-site Gibbs sampling method. The line corresponds to a one to one relationship of inferred and true BV.

3 MB of RAM. Further sensitivity analysis is shown in Appendix S3. The impact of prior assumptions on estimated posterior distributions is shown in Fig. S1. Results of the analysis of the second simulated data set are shown in Appendix S1.

Table 2 shows effective sample sizes (ESS; Kass *et al.* 1998), that is, sample size adjusted for autocorrelation and autocorrelations of the obtained MCMC using the standard Gibbs samplers. The difference in ESS and autocorrelation was dependent on the pedigree size, where the smallest pedigree resulted in the best mixing and lowest level of autocorrelation between samples in the MCMC. The MCMC analysis of data set 1, however, resulted in low ESS and considerable autocorrelations. The convergence statistics was calculated within the R packages CODA (Plummer *et al.* 2006) and boa (Smith 2007).

ANALYSED WHEAT DATA 2

Summary statistics of inferred posterior distributions for h^2 and BV on the analysed data set 2 is shown in Table 3. The performances of the analytic and the standard MCMC approaches were similar in terms of accuracy of inferred parameters. The posterior mean of h^2 obtained with the analytic method was, however, slightly higher than corresponding point estimates obtained with MCMC and REML. On the other hand, the K–L divergence between inferred posteriors of h^2 was very low, for example close to zero. Furthermore, the correlation between posterior mean of BV obtained by analytic and MCMC methods was practically one. Worth noting is the relatively poor mixing and low ESS of the chains obtained with the standard methods, both for single-site and blocked sampling, as seen in Table 2. The standard blocked MCMC sampler required about 30–60 and three times more computational time compared to the time required by the analytic and single-site Gibbs sampler, respectively.

A low number of bins (i.e. 20) was needed to obtain the same accuracy as with, for example, 1000 bins. As a result, the required computational time of the analytic method was much less: up to 100-fold of the computational time required by the standard MCMC. Accurate point estimates of both h^2 and BV are obtained after approximately 6 s, which was even faster than the REML method. The precision was, however, reduced in the 20 bin case due to the large bin size (i.e. wider 95% credible region). To infer point estimates of BVs, the computational time required by the CE approach was much less, when the number of bins were either 20 or 100, than required by the blMCMC approach. The opposite was found when analysing 1000 bins: the blMCMC approach outperformed the CE approach. An extensive sensitivity analysis is shown in

Table 2. Markov Chain Monte Carlo (MCMC) autocorrelation

Method	Data	Pedigree size	Missing values	ESS	Lag 1	Lag 5	Lag 10	Lag 50
MCMC1	1	5865	1200	685·1	0·804	0·594	0·458	0·088
MCMC2	1	5865	1200	434·2	0·891	0·623	0·431	0·002
MCMC1	2	599	0	924·5	0·643	0·388	0·286	0·073
MCMC2	2	599	0	641·0	0·844	0·514	0·283	−0·012
MCMC1	3	700	0	3255·5	0·549	0·206	0·085	−0·001
MCMC2	3	700	0	1523·1	0·722	0·231	0·063	−0·011

Statistics on MCMC convergence of the analysed pedigrees where ESS is the effective sample size, Lag is the time lag of the thinned MCMC chain. Single-site updating method is denoted MCMC1, while blocked sampling method is denoted MCMC2. In total, 225 000 iterations were simulated in each chain where the first 25 000 were discarded and every 10th saved, leaving the size of the chain to 20 000. For MCMC2, 25 000 iterations were simulated where the first 5000 iterations were discarded and every 2nd saved leaving the size of the chain to 10 000.

Table 3. Analysed data 2

		Heritability					Correlations		Computational time				
Model	nbins	Mode	Mean	SD	95% CI	K–L	$cor(\mathrm{BV}_A, \mathrm{BV}_{\mathrm{SM}})$	$cor(\mathrm{BV}_A, \mathrm{BV}_{\mathrm{ML}})$	Δt_C	Δt_M	t_{h^2}	$t_{\mathrm{BV},C}$	$t_{\mathrm{BV},M}$
Analytic	20	0·333	0·351	0·059	[0·091, 0·474]	0·061	1·000	1·000	0·11	0·94	0·04	0·07	0·90
Analytic	100	0·333	0·351	0·059	[0·231, 0·462]	0·045	1·000	1·000	0·39	1·04	0·12	0·27	0·92
Analytic	1000	0·341	0·350	0·059	[0·232, 0·462]	0·050	1·000	1·000	3·62	2·26	0·94	2·68	1·32
MCMC1	–	0·330	0·333	0·059	[0·216, 0·447]	–	–	–	18·06	–	–	–	–
MCMC2	–	0·342	0·335	0·060	[0·216, 0·454]	–	–	–	57·12	–	–	–	–
REML	–	0·339	–	–	–	–	–	–	–	–	–	–	–

Summary statistics of inferred h^2 obtained from analysis of Wheat data set 2, where nbins is the number of bins of h^2, SD is the standard deviation, CI is the credible interval, K–L is the Kullback–Leibler divergence of inferred posteriors of h^2 obtained from analytic and single-site MCMC methods. Single-site updating method is denoted MCMC1, while blocked sampling method is denoted MCMC2. Correlation between inferred breeding values (BV) using the analytic approach and single-site Gibbs sampler and restricted maximum likelihood (REML) is denoted $cor(\mathrm{BV}_A, \mathrm{BV}_{\mathrm{SM}})$ and $cor(\mathrm{BV}_A, \mathrm{BV}_{\mathrm{ML}})$, respectively, for both conditional expectations (CE) and blMCMC inference methods, as both methods resulted in equal correlations. The total computational time for the analytic approach with either CE or blMCMC is denoted Δt_C and Δt_M, respectively. The computational time for the heritability estimation, CE and blMCMC to obtain inferred BVs is denoted t_{h^2}, $t_{\mathrm{BV},C}$ and $t_{\mathrm{BV},M}$. All time units are given in minutes. The parameter range of h^2 is between 0 and 0·8.

Appendix S3. The impact of prior assumptions on estimated posterior distributions is shown in Fig. S2. A two random effects case with additive and dominance genetic effects is shown in Appendix S4.

Discussion

New sequencing techniques allows obtaining genomewide, dense marker maps for not only model species but also species of interest to ecologists, conservation geneticists and breeders. The amount of data is expected to increase rapidly in the near future which, in turn, will require efficient and powerful statistical inference methods in order to facilitate learning of parameters of interest. We have presented a novel Bayesian approach for analysing complex traits and drawing inferences in animal models. There are two major advantages of the proposed approach compared to standard MCMC approaches. First, the computational burden is reduced, sometimes considerably so, on the data analysed here. Second, convergence is not a concern here, which sometimes can be very problematic in standard MCMC, in particular, if single-site updating of the parameters is performed. Obtained results are very encouraging: we obtain practically identical results as obtained with standard MCMC and REML approaches over two simulated example pedigrees and a Wheat pedigree with dense marker maps available. Sensitivity analyses suggest that the proposed method is robust to various prior assumptions on the inverse lambda, which is proportional to the heritability. The straight forward way to incorporate prior information on the heritability, either indirectly via λ_u^{-1} or directly, highlights the benefit of the approach over REML and, to some extent, over standard MCMC approaches. Such prior information could for example be obtained from a meta-analysis for the trait and species under consideration.

It should be pointed out, though, that we used two standard MCMC sampling implementations, via the Gibbs sampler, as a reference samplers, which is often utilized for drawing inferences in animal models (Sorensen & Gianola 2002). However, there exist more efficient MCMC implementation methods that reduce the computational burden by avoiding searching the entire parameter space. These adaptive MCMC methods make use of gradient information to propose a new parameter proposal density, as for example the Langevin–Hastings algorithm (Roberts & Tweedie 1996). In a comparative study on the efficiency of various MCMC updating strategies on three real pedigrees, Waagepetersen, Ibánêz-Escriche & Sorensen (2008) found that Langevin–Hastings reduced computational time and suggested a joint Langevin–Hastings and normal approximation scheme based on Taylor expansion for both saved computing time and maintaining small autocorrelations throughout the estimation procedure. Other, related approaches have been proposed that approximate the posterior distribution using the Laplace approximation method and, thereby, avoid using MCMC simulation (Hofer & Ducrocq 1997). See also the suggested method by Steinsland & Jensen (2010) and a recent, user-friendly implementation of the Laplace approximation method: the animal-INLA package (Holand *et al.* 2013). One drawback with the Laplace approximation method, compared to ours, is that if the posterior distribution is multimodal, maximization procedure will find only a single mode and approximation with a normal distribution might severely bias credible regions and result in erroneous inference. Similar problems arises with classic REML methods for unidentifiable likelihood functions.

Another major advantage with our approach in terms of computational efficiency has not been utilized here, but is likely to increase the efficiency of the approach. Each bin calculation of the likelihood is independent, so that an analyst could, for example, infer lambda inverse in which order of bins one prefers, as opposed to MCMC where the parameter state in one iteration is dependent on the state in the previous iteration, as parameters are drawn from conditional posterior distributions. The advantage, from a computational efficiency point of view, is that the analysis could be parallelized on a multiple core computer so that bin calculations are divided and executed on separate threads. In doing so, the total computational time required could be further reduced, probably considerably so, depending on the hardware available and the size of the analysed pedigree. This multicore computing procedure is often proposed for regular Monte Carlo or resampling simulations, where a large number of independent iterations need to be executed.

A major issue with using regular MCMC methods is identifying when convergence of the chain is reached and how many samples are needed to ensure drawing from the stationary conditional posterior distribution. High conditional posterior correlations might introduce heavy dependencies in the chain which results in poor mixing. Typically, a Gibbs sampler might get stuck in a small subspace of the entire parameter space for a large number of iterations. As a result, massive computational efforts are needed to reduce MC errors to acceptable levels and to obtain a good estimation of the marginal posterior distributions of all parameters of interest. This is a particular issue with the single-site Gibbs sampler, implemented as reference sampler here, where high levels of autocorrelation and low effective sample size were obtained. A blocked implementation of the Gibbs sampler, also implemented as a reference sampler here, where all parameters in the model are updated jointly (Garcia-Cortes & Sorensen 1996), has been shown to improve mixing and reduce the autocorrelations. On the other hand, the blocked sampler tends to be computationally expensive as the large linear system of equations needs to be repeatedly solved. In order to implement an efficient Gibbs sampler, both in terms of computational speed and mixing properties, Waldmann *et al.* (2008) combined the single-site and blocked samplers into a hybrid sampler and reparameterized the random additive and dominance polygenic effects. Although the resulting sampler reduced the computational time compared to a pure block sampler and improved the mixing property compared to the single-site sampler, the required computational effort was still massive. By our approach introduced here, the convergence and mixing problems are avoided.

In an animal model framework, the inverse of the realized relationship matrix, $\mathbf{A}_G$, is needed in order to infer heritability and BV. In practice, obtaining the inverse may not be feasible due to introduction of dependencies among columns in $\mathbf{A}_G$ (i.e. multicolinearity) which causes non-positive definiteness. This problem might arise, for example, if $\mathbf{A}_G$ has been calculated based on too few markers, if clones or monozygotic twins are present in the pedigree, the choice of allele coding and if dependencies of marker profiles are present (Frentiu *et al.* 2008; VanRaden 2008; Strandén & Christensen 2011; Piepho *et al.* 2012). The method proposed here does not need $\mathbf{A}_G$ to be positive definite, as diagonal elements are added to calculate $\Sigma^\star$. Thus, the heritability can be learned based on the exact $\mathbf{A}_G$, and no *ad hoc* methods are needed to make $\mathbf{A}_G$ invertible. We investigated the impact of using the exact $\mathbf{A}_G$ on estimated posterior of h^2 compared to results obtained with the modified $\mathbf{A}_G^\star$ and found conflicting pattern: in the analysed data set 1, point estimates of h^2 agreed closely, whereas a large discrepancy of point estimates were found in the analysed Wheat data set. These results might reflect the difference in population size and marker coverage of the analysed data, influencing the outcome of the relationship estimator used here, which has been proposed by Frentiu *et al.* (2008) and Sillanpää (2011). Hence, the problem of non-positive definiteness of the covariance matrix might be more important in applications where the coverage of the marker map is not perfect and the size of the analysed pedigree is small. Further test are needed to examine the impact of marker density and pedigree size, preferably by analysing simulated data with known parameter values. Another possible explanation to the obtained differences in inferred parameters might be the use of the relationship estimator in the artificial Wheat population. As the population consists of variety lines, a deficiency of heterozygotes could bias estimated relationships.

It is common, in animal model applications, that multiple random terms are included in the linear model. For example, maternal effects arise when the phenotype of the mother influences the phenotype of her offspring in addition to the additive effect and non-additive genetic effects which introduces nonlinear dependency between phenotypes and genotypes due to the interactions within and between loci (Lynch & Walsh 1998; Hallander & Waldmann 2007). Typically, these additional effects are efficiently modelled within the animal model framework as random effects (e.g. Lynch & Walsh 1998; Sorensen & Gianola 2002; Kruuk 2004). In Appendix S4, we have shown how two random effects can efficiently be handled in the proposed approach to infer the joint posterior distribution of two lambda parameters, proportional to the heritability and the dominance genetic proportion, respectively. Although we did not include the breeding value inference step, the computational efficiency of our approach in the two random effects case seems encouraging. Further improvements in computational efficiency could involve discretizing the parameter space of λ in two steps: one initial analysis where few bins are utilized and a second analysis where new bins are introduced near the bin having maximum posterior value. For multiple random components, efficient search algorithms, such as the simulated annealing technique, could help to find maximum posterior value of each λ. Furthermore, for models with a large number of location parameters, and particular for low to medium number of phenotypic observations, a well-known problem of MCMC inference is parameter identifiability and high posterior correlation among inferred parameters (Sorensen & Gianola 2002; Gelman *et al.* 2004; Waldmann *et al.* 2008). In such situation, our approach will benefit from avoiding convergence problems which, in turn, can result in more accurate and robust learning of genetic parameters and reduced computational time. To handle multiple random effects with our suggested approach and extend analysis to inference of BV (and the additional location effects) needs further investigation in the future.

Acknowledgements

This work was initialized when J.A. (formerly Hallander) was a post-doc at the Swedish University of Agricultural Sciences. The associate editor and reviewers are acknowledged for suggestions that improved earlier versions of the paper. In addition, we wish to acknowledge Petri Koistinen, University of Helsinki, for useful discussions. This work was supported by funding from the Research School of Tree Breeding and Forest Genetics for J.A. and research grants from the Academy of Finland and the University of Helsinki's Research Funds for M.J.S. We have no conflict of interest to declare.

References

Aulchenko, Y.S., de Koning, D.J. & Haley, C. (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis. *Genetics*, **177**, 577–585.

Bauer, A.M., Hoti, F., Reetz, T.C., Schuh, W.-D., Leon, J. & Sillanpää, M.J. (2009) Bayesian prediction of breeding values by accounting for genotype-by-environment interaction in self-pollinating crops. *Genetics Research*, **91**, 193–207.

Brommer, J.E., Rattiste, K. & Wilson, A.J. (2008) Exploring plasticity in the wild: laying date temperature reaction norms in the common gull *Larus canus*. *Proceedings of the Royal Society of London, Series B*, **275**, 687–693.

Crossa, J., de los Campos, G., Perez, P., Gianola, D., Burgueno, J., Araus, J.L. *et al.* (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, **186**, 713–724.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B. & Woolliams, J.A. (2010) The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, **185**, 1021–1031.

Frentiu, F.D., Clegg, S.M., Chittock, J., Burke, T., Blows, M.W. & Owens, I.P. (2008) Pedigree-free animal models: the relatedness matrix reloaded. *Proceedings of the Royal Society of London, Series B*, **275**, 639–647.

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

Gasbarra, D., Pirinen, M., Sillanpää, M.J. & Arjas, E. (2009) Bayesian quantitative trait locus mapping based on reconstruction of recent genetic histories. *Genetics*, **183**, 709–721.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004) *Bayesian Data Analysis*, 2nd edn. Chapman and Hall/CRC, New York.

Gianola, D. & van Kaam, J.B.C.H.M. (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, **178**, 2289–2303.

Gilks, W.R., Spiegelhalter, D.J. & Richardson, S. (1995) *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, London.

Golan, D. & Rosset, S. (2011) Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics*, **27**, 1317–1323.

Golub, G. & van Loan, C. (1996) *Matrix Computations*, 3rd edn. The Johns Hopkins University Press, London.

Hallander, J. & Waldmann, P. (2007) The effect of non-additive genetic interactions on selection in multi-locus genetic models. *Heredity*, **98**, 349–359.

Hallander, J., Waldmann, P., Wang, C. & Sillanpää, M.J. (2010) Bayesian inference of genetic parameters based on conditional decompositions of multivariate normal distributions. *Genetics*, **185**, 645–654.

Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–447.

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

Hofer, A. & Ducrocq, V. (1997) Computing marginal posterior densities of genetic parameters of a multiple trait animal model using Laplace approximation or Gibbs sampling. *Genetics Selection Evolution*, **29**, 427–450.

Holand, A.M., Steinsland, I., Martino, S. & Jensen, H. (2013) *Animal Models and Integrated Nested Laplace Approximations. G3 (Bethesda)*, **3**, 1241–1251.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Kang, H.M., Hoon-Sul, J., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimen, N.B., Sabatti, C. & Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**, 348–354.

Kass, R.E. & Raftery, A.E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kass, R.E., Carlin, B.P., Gelman, A., Neal, R. (1998) Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician*, **52**, 93–100.

Kruuk, L.E.B. (2004) Estimating genetic parameters in natural populations using the 'animal model'. *Philosophical Transactions of the Royal Society of London, Series B*, **359**, 873–890.

Kullback, S. & Leibler, R.A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

Legarra, A. & Misztal, I. (2008) Computing strategies in genome-wide selection. *Journal of Dairy Science*, **91**, 360–366.

Lund, M.S., Sahana, G., de Koning, D.J., Su, G. & Carlborg, O. (2009) Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. *BMC Proceedings*, **3**, 1.

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

Maenhout, S., DeBaets, B. & Haensert, G. (2009). Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. *Theoretical and Applied Genetics*, **118**, 1181–1192.

Mathew, B., Bauer, A.M., Koistinen, P., Reetz, T.C., Leon, J. & Sillanpää, M.J. (2012) Bayesian adaptive Markov chain Monte Carlo estimation of genetic parameters. *Heredity*, **109**, 235–245.

Meuwissen, T.H.E. & Goddard, M.E. (2010) Accurate prediction of genetic values for complex traits by whole genome resequencing. *Genetics*, **185**, 623–631.

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.

Meyer, K. (2007) WOMBAT A tool for mixed model analyses in quantitative genetics by REML. *Journal of Zhejiang University Science B*, **8**, 815–821.

O'Hagan, A. & Forster, J.J. (2004) *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*, 2nd edn. Arnold, London, UK.

Pemberton, J.M. (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society of London, Series B*, **275**, 613–621.

Piepho, H.P. (2009) Ridge regression and extensions for genomewide selection in maize. *Crop Science*, **49**, 1165–1176.

Piepho, H.P., Ogutu, J.O., Schulz-Streeck, T., Estaghvirou, B., Gordillo, A. & Technow, F. (2012) Efficient computation of ridge-regression BLUP in genomic selection in plant breeding. *Crop Science*, **52**, 1093–1104.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006) CODA: convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11.

Resende, M.F.R. Jr, Munoz, P., Resende, M.D.V., Garrick, D.J., Fernando, R.L., Davis, J.M. *et al.* (2012) Accuracy of genomic selection methods in a standard data set of Loblolly pine (*Pinus taeda* L.). *Genetics*, **190**, 1503–1510.

Riester, M., Stadler, P.F. & Klemm, K. (2009) Reconstruction of wild multi-generation pedigrees. *Bioinformatics*, **25**, 2134–2139.

Ritland, K. (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology*, **9**, 1195–1204.

Roberts, G.O. & Tweedie, R.L. (1996) Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli*, **2**, 341–363.

Santure, A.W., Stapley, J., Ball, A.D., Birkhead, T.R., Burke, T. & Slate, J. (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, **19**, 1439–1451.

Searle, S.R., Casella, G. & McCulloch, C.E. (1992) *Variance Components*. John Wiley & Sons, New York.

Sillanpää, M.J. (2011) On statistical methods for estimating heritability in wild populations. *Molecular Ecology*, **20**, 1324–1332.

Smith, B.J. (2007) boa: an R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, **21**, 1–37.

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

Steinsland, I. & Jensen, H. (2010) Utilizing Gaussian Markov Random Field properties of Bayesian animal models. *Biometrics*, **66**, 763–771.

Strandén, I. & Christensen, O.F. (2011) Allele coding in genomic evaluation. *Genetics Selection Evolution*, **43**, 25.

Thompson, R. (2008) Estimation of quantitative genetic parameters. *Proceedings of the Royal Society of London, Series B*, **275**, 679–686.

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science*, **91**, 4414–4423.

Visscher, P.M., Hill, W.G. & Wray, N.R. (2008) Heritability in the genomics era: concepts and misconceptions. *Nature Reviews, Genetics*, **9**, 255–266.

Waagepetersen, R., Ibánêz-Escriche, N. & Sorensen, D. (2008) A comparison of strategies for Markov chain Monte Carlo computation in quantitative genetics. *Genetics Selection Evolution*, **40**, 161–176.

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.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S. Henders, A.K. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**, 565–569.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203–208.

Received 19 June 2013; accepted 15 July 2013
Handling Editor: Jarod Hadfield

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Apendix S1. Additional analyzed synthetic data.

Apendix S2. Model selection for covariance structures.

Apendix S3. Sensitivity analysis.

Apendix S4. The two random effects case.