BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits

Takeshi Hayashi[1*] and Hiroyoshi Iwata[2]

Abstract

Background: Genomic selection is an effective tool for animal and plant breeding, allowing effective individual selection without phenotypic records through the prediction of genomic breeding value (GBV). To date, genomic selection has focused on a single trait. However, actual breeding often targets multiple correlated traits, and, therefore, joint analysis taking into consideration the correlation between traits, which might result in more accurate GBV prediction than analyzing each trait separately, is suitable for multi-trait genomic selection. This would require an extension of the prediction model for single-trait GBV to multi-trait case. As the computational burden of multi-trait analysis is even higher than that of single-trait analysis, an effective computational method for constructing a multi-trait prediction model is also needed.

Results: We described a Bayesian regression model incorporating variable selection for jointly predicting GBVs of multiple traits and devised both an MCMC iteration and variational approximation for Bayesian estimation of parameters in this multi-trait model. The proposed Bayesian procedures with MCMC iteration and variational approximation were referred to as MCBayes and varBayes, respectively. Using simulated datasets of SNP genotypes and phenotypes for three traits with high and low heritabilities, we compared the accuracy in predicting GBVs between multi-trait and single-trait analyses as well as between MCBayes and varBayes. The results showed that, compared to single-trait analysis, multi-trait analysis enabled much more accurate GBV prediction for low-heritability traits correlated with high-heritability traits, by utilizing the correlation structure between traits, while the prediction accuracy for uncorrelated low-heritability traits was comparable or less with multi-trait analysis in comparison with single-trait analysis depending on the setting for prior probability that a SNP has zero effect. Although the prediction accuracy with varBayes was generally lower than with MCBayes, the loss in accuracy was slight. The computational time was greatly reduced with varBayes.

Conclusions: In genomic selection for multiple correlated traits, multi-trait analysis was more beneficial than single-trait analysis and varBayes was much advantageous over MCBayes in computational time, which would outweigh the loss of prediction accuracy caused by the approximation procedure, and is thus considered a practical method of choice.

Keywords: Genomic selection, Multiple traits, Bayesian regression, MCMC iteration, Variational approximation

* Correspondence: hayatk@affrc.go.jp
[1]Agroinformatics Division, National Agriculture and Food Research Organization, Agricultural Research Center, Kannondai, Tsukuba, Ibaraki 305-8666, Japan
Full list of author information is available at the end of the article

Background

A huge number of genome-wide polymorphisms have recently been elucidated in livestock and crops with the development of sequencing technologies. High-throughput genotyping systems, such as high-density SNP chips containing several tens or hundreds of thousands of genome-wide SNP markers and GBS (genotyping by sequence) [1], have become available to efficiently identify genotypes of individuals for a large number of SNPs at low cost. Consequently, genomic selection [2] is attracting attention as a new breeding technology utilizing the information of genome-wide dense SNP markers. In genomic selection, a model to predict genomic breeding value (GBV) based on genome-wide SNP genotype is firstly constructed using a training population consisting of individuals with both SNP genotypes and phenotypes of a trait and, subsequently, using this model, the GBVs of a trait are predicted for individuals in the tested population, which are selection candidates, based on their SNP genotypes, allowing effective individual selection to be performed without phenotypic records using the predicted GBV. Therefore, genomic selection requires the construction of prediction models being able to accurately relate genotypes of genome-wide SNPs to GBV.

In the original study of genomic selection by Meuwissen et al. [2], two Bayesian methods were presented for construction of prediction models, which were referred to as BaeysA and BayesB and have been used for the studies of genomic selection in their original or modified forms [3,4]. The model of BayesA is equivalent to the Bayesian shrinkage regression (BSR) model [5], where all SNPs are included in the prediction model as covariates and the estimates of SNP effects are shrunk by assuming a normal distribution with mean 0 and SNP-specific variance as prior distributions of SNP effects, resulting in negligible estimates for the effects of many SNPs irrelevant to phenotype, but significantly large estimates for the effects of SNPs contributing to phenotype. On the other hand, the BayesB method is regarded as a variant of Bayesian stochastic search variable selection (BSSVS) [6], where the prior probability, π, that a SNP has zero effect and is removed from the model was considered in the model fitting to obtain the best model explaining the phenotypes with a small number of SNP effects. A normal distribution with mean 0 and SNP-specific variance is assumed for the prior distribution of SNP effects in BayesB, as in BayesA, if SNPs are included in the model with probability 1-π and, otherwise, both the mean and variance are fixed at 0 with probability π.

In BayesA and BayesB, an additional hierarchical structure is induced for this SNP-specific variance, where an inverted chi-square distribution with degree of freedom ν and scale parameter S is adopted as the prior distribution of the variance. However, only the information of the relevant single SNP can be used for the posterior inference of this SNP-specific variance regardless of the number of genotypes or phenotypes. Due to the insufficient Bayesian learning for the variance, the degree of shrinkage for the estimates of SNP effects is largely influenced by the prior setting for ν and S in BayesA and BayesB [7]. In BaeysB, the sparseness of the model and the prediction accuracy are also greatly affected by the prior probability, π, that a SNP has zero effect, which is treated as a known fixed value. To overcome these drawbacks of BayesA and BayesB, the modified methods such as BayesC and BayesD were proposed. In BayesC, a single variance common to all SNPs is adopted for the prior distribution of SNP effects while BayesD allows S to be inferred from the data, given ν. Furthermore, the step for inferring π, which is given as a fixed value in BayesB, can be incorporated in the procedure of BayesC and BayesD and the corresponding methods are termed BayesCπ and BayesDπ, respectively [4].

These Bayesian methods were mainly developed for genomic selection of a single trait. However, actual breeding of animals and plants often aims to simultaneously improve multiple correlated traits. Therefore, joint prediction of GBVs for multiple traits, taking into consideration the correlation structure between traits, is suitable for multi-trait genomic selection, which requires the extension of existing methods for single-trait GBV prediction to multi-trait case. In QTL mapping methods that use similar models to those of genomic selection, some researchers have developed multi-trait models [8-10]. Xu et al. [8] extended the BSR model to a multi-trait case introducing the correlation structure between traits in estimation of QTL effects and other non-genetic effects. Banerjee et al. [9] employed a Bayesian composite model space approach [11,12] for multi-trait QTL mapping, where a variable indicating inclusion or exclusion of each QTL was incorporated for constructing a model with the smallest possible number of QTLs, a similar approach to BSSVS adopted in BayesB, and each trait was allowed to have a different model by assuming the trait-specific effects for QTLs and non-genetic factors that were assumed to be uncorrelated between traits. Meuwissen and Goddard [10] proposed a multi-trait BSSVS model for QTL mapping. These methods can be applied for prediction of multi-trait GBVs in genomic selection. Calus and Veerkamp [13] applied the method proposed in [10] with some modification for multi-trait GBV prediction to compare the accuracy between single-trait prediction and multi-trait prediction in genomic selection.

The computational procedure with MCMC iteration is generally used for the Bayesian methods to estimate parameters in the models, which become complicated models in genomic selection, including a huge number of SNPs as covariates and SNP effects that are estimated as their regression coefficients. The computational burden of MCMC-based Bayesian methods, which requires

a long time until convergence when estimating many parameters is huge even in single-trait GBV prediction and would be further increased in the case of multiple traits, thus hindering the MCMC procedure depending on the number of traits to be jointly analyzed. Therefore, it would be necessary to devise a solution for reducing the computational burden of Bayesian methods for multi-trait GBV prediction. So far, some non-MCMC computational procedures for Bayesian methods have been proposed in QTL mapping and genome-wide association study, including EM-algorithm [14] and variational approximation [15,16]. The EM-algorithm was also applied to single-trait GBV prediction, successfully reducing the computational time [17,18].

In this paper, we propose Bayesian methods for multi-trait GBV prediction, in which BSR models allowing variable selection are developed and MCMC procedures for estimating model parameters including SNP effects are described as well as a computationally cost-effective non-MCMC method using variational approximation as an alternative computational procedure. Hereafter, the Bayesian methods based on MCMC iteration and variational approximation are referred to as MCBayes and varBayes, respectively. The multi-trait Bayesian models described here include a Bayesian shrinkage regression (BSR) models that are equivalent to those adopted by BayesA and BayesD when variable selection is not conducted, and the models of BSSVS that are regarded as slightly modified versions of BayesB and BayesDπ methods when a step of variable selection step is incorporated. We develop computationally-effective variational approximation procedures to construct the posterior distributions of parameters of these Bayesian models.

Using simulated datasets consisting of genotypes of genome-wide dense SNPs and phenotypes of three correlated traits with high and low heritabilities, we investigated the differences in prediction accuracy for each trait between multi-trait analysis, where GBVs of three traits were simultaneously predicted taking the correlation structure between traits into consideration, and single-trait analysis, where each trait was separately predicted for GBV. We also evaluated the prediction accuracy of the varBayes methods in comparison with MCBayes. Moreover, we investigated the performance of multi-trait analysis in simulated data including missing phenotypes.

Methods

In this section, we describe Bayesian models for multi-trait GBV prediction and computational procedures for Bayesian estimation of the model parameters, including construction of posterior distributions of the parameters using MCMC iteration and variational approximation. Here, we consider the statistical models for BSR and BSSVS, which are shown to be equivalent to BayesD and similar to BayesDπ, respectively. In these Bayesian models concerned, BSR is regarded as a special version of BSSVS by setting π =0, where π is a prior probability that a SNP does not contribute to traits; thus, we present the multi-trait GBV prediction models in a unified fashion in terms of BSSVS.

We assume that the number of SNPs genotyped is N and a training dataset including n individuals with phenotypic records of multiple traits, where the number of traits is denoted by T, and SNP genotypes is available for estimating parameters in the prediction model. We also assume that a test dataset consists of individuals with only SNP genotypes, for each of which GBV is predicted. We denote two alleles at each SNP by 0 and 1 and three genotypes by '0_0', '0_1', and '1_1'.

Models for Bayesian stochastic search variable selection in multi-trait genomic selection

We propose the following Bayesian multi-locus linear model for the phenotypes of T traits of the ith individual, which are denoted as a Tx1 vector $\mathbf{y}_i$ (i = 1,2,. . .,n), as a BSSVS model for multi-trait GBV prediction:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \sum_{l=1}^{N} \gamma_l u_{il}\mathbf{g}_l + \mathbf{e}_i, \tag{1}$$

where $\mathbf{b}$ is a $f \times 1$ vector of non-genetic effects including interceptions of the model with $\mathbf{X}_i$ being a $T \times f$ design matrix linking $\mathbf{b}$ to the ith individual; u_{il} is a variable indicating the genotype of the ith individual at the lth SNP taking a value of −1, 0 or 1 corresponding to the genotypes, '0_0', '0_1', or '1_1', respectively; $\mathbf{g}_l$ is a $T \times 1$ vector of the effects of the lth SNP on the phenotypes of T traits; γ_l is a variable indicating the inclusion of the lth SNP in the model with 1 or 0 depending on whether or not the lth SNP is included in the model; and $\mathbf{e}_i$ is a $T \times 1$ vector of residual errors following a T-variate normal distribution $N(\mathbf{0}, \mathbf{\Sigma}_e)$ with $\mathbf{0}$ being a $T \times 1$ vector of zeros and $\mathbf{\Sigma}_e$ being a TxT covariance matrix of $\mathbf{e}_i$.

Within the Bayesian framework, prior distributions are assigned to the parameters of the model (1). We assume that the priors of the elements of $\mathbf{b}$ are the improper uniform distribution over the possible values. The prior probabilities that γ_l =1 and γ_l =0 are presented as 1- π and π, respectively, considering the prior probability that a SNP does not contribute to the trait and is excluded from the model is π. The prior distribution of $\mathbf{g}_l$ given γ_l has the form

$$\mathbf{g}_l|\gamma_l \sim \begin{cases} N(\mathbf{0}, \mathbf{\Sigma}_{gl}) & (\gamma_l = 1) \\ \delta(\mathbf{0}) & (\gamma_l = 0) \end{cases}, \tag{2}$$

where $\mathbf{\Sigma}_{gl}$ is a $T \times T$ matrix that is the variance and covariance matrix of $\mathbf{g}_l$ given γ_l =1 and δ ($\mathbf{0}$) is the delta

function that concentrates a total mass at zero for all T elements of $\mathbf{g}_l$. We induce a hierarchical structure for $\mathbf{\Sigma}_{gl}$ by assigning an inverse Wishart distribution with degree of freedom ν and scale parameter $\mathbf{S}$, denoted by $\mathrm{IW}_T(\nu,\mathbf{S})$, to the prior distribution of $\mathbf{\Sigma}_{gl}$:

$$\Sigma_{gl}\sim \mathrm{IW}_T(\nu,\mathbf{S})\propto \left|\mathbf{S}\right|^{\nu/2}\left|\mathbf{\Sigma}_{gl}\right|^{-(\nu+\mathrm{T}+1)/2}\exp\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{S}\mathbf{\Sigma}_{gl}{}^{-1}\right)\} \tag{3}$$

Although the SNP effect $\mathbf{g}_l$ is zero and irrelevant to $\mathbf{\Sigma}_{gl}$ when γ_l =0 as shown in (2), we assume that $\mathbf{\Sigma}_{gl}$ is a priori distributed as given in (3) regardless of γ_l. We treat ν as a given fixed value but infer $\mathbf{S}$ from the data within the Bayesian framework. For simplicity, we assume here that $\mathbf{S}$ is a diagonal matrix with the kth diagonal element being s_k , that is $\mathbf{S}=\mathrm{diag}(s_k)$ $(k=1,2,\ldots,T)$, and we adopt the improper uniform distribution over positive values for a prior distribution of s_k. The residual variance and covariance matrix $\mathbf{\Sigma}_e$ is assumed to have a uniform distribution over the positive definite matrices as a prior distribution.

Denoting these parameters of the Bayesian model collectively by $\boldsymbol{\theta}$, the prior distribution of $\boldsymbol{\theta}$ by $p(\boldsymbol{\theta})$ and observed phenotypes $\mathbf{y}_i$ and SNP genotypes u_{il} $(i=1,2,\ldots,n;l=1,2,\ldots,N)$ by $\mathbf{Y}$ and $\mathbf{U}$, respectively, we can write the posterior distribution of $\boldsymbol{\theta}$, $g(\boldsymbol{\theta}|\ \nu,\mathbf{Y},\mathbf{U})$, as

$$\begin{aligned} g(\theta|\nu,\mathbf{Y},\mathbf{U}) &\propto |\mathbf{\Sigma}_{\mathrm{e}}|^{-\boldsymbol{n}/2}\exp(-\frac{1}{2}\sum_{\mathrm{i}=1}^{\boldsymbol{n}}\mathbf{y}_{\mathrm{i}}*'\mathbf{\Sigma}_{\mathrm{e}}{}^{-1}\mathbf{y}_{\mathrm{i}}*)\\ &\times\prod_{l=1}^{N}[\{(1-\pi)|\mathbf{\Sigma}_{gl}|^{-1/2}\exp(-\frac{1}{2}\mathbf{g}_l{}'\mathbf{\Sigma}_{gl}{}^{-1}\mathbf{g}_l)\}^{\gamma_l}\{\pi\delta(\mathbf{0})\}^{1-\gamma_l}]\\ &\times\prod_{l=1}^{N}\left|\mathbf{S}\right|^{\nu/2}\left|\mathbf{\Sigma}_{gl}\right|^{-(\nu+T+1)/2}\exp(-\frac{1}{2}\mathrm{tr}\mathbf{S}\mathbf{\Sigma}_{gl}{}^{-1}) \end{aligned} \tag{4}$$

from (1), (2) and (3), where $\mathbf{y}_i{}^*$ is a residual given by $\mathbf{y}_i*=\mathbf{y}_i-\mathbf{X}_i\mathbf{b}-\sum_{l=1}^{N}\gamma_l u_{il}\mathbf{g}_l$.

Given that $\mathbf{S}$ is fixed, posterior distribution (4) is equivalent to that of BayesA extended to multi-trait case when $\pi=0$ leading to γ_l =1 for all SNPs. When $\pi>0$, the selection of SNPs to be included in the model is performed as in BayesB. However, it should be noted that posterior distribution (4) is not equivalent to that of the multi-trait version of BayesB, which supposes that $\mathbf{\Sigma}_{gl}$ is a zero matrix as well as $\mathbf{g}_l$ when γ_l =0, while, in (4), the prior distribution of $\mathbf{\Sigma}_{gl}$ is assumed to be $\mathrm{IW}_T(\nu,\mathbf{S})$ regardless of γ_l. The form of posterior distribution (4) allows Gibbs sampling in the MCMC estimation for all parameters including $\mathbf{g}_l$ and $\mathbf{\Sigma}_{gl}$. The full conditional posterior distributions of parameters used in MCBayes are described in Additional file 1.

Variational approximation procedure for multi-trait Bayesian model

We adopted variational approximation as an alternative to MCMC iteration for constructing marginal posterior distributions of parameters based on joint posterior distribution (4). In the variational approximation procedure, the joint posterior distribution is approximated by the product of functions for subsets of parameters with lower dimension. Briefly, we assume that $g(\boldsymbol{\theta}|\ \nu,\mathbf{Y},\mathbf{U})$ is approximated by a function of $\boldsymbol{\theta}$, $q(\boldsymbol{\theta})$, which is factorized as $q(\boldsymbol{\theta})=q_1(\boldsymbol{\theta}_1)\ q_2(\boldsymbol{\theta}_2)\ldots q_K(\boldsymbol{\theta}_K)$, where $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_K$ are mutually exclusive subsets of $\boldsymbol{\theta}$ such that $\cup_{k=1}^{K}\boldsymbol{\theta}_k=\boldsymbol{\theta}$ and q_k $(\boldsymbol{\theta}_k)$ $(k=1,2,\ldots,K)$ may generally depend on $\nu,\mathbf{Y}$ and $\mathbf{U}$ although this dependence is omitted here for simplicity. This approximating marginal posterior distribution, $q_k(\boldsymbol{\theta}_k)$, is referred to as the variational posterior of $\boldsymbol{\theta}_k$ [19]. The form of $q_k(\boldsymbol{\theta}_k)$ is determined such that the Kullback–Leibler divergence between $g(\boldsymbol{\theta}|\nu,\mathbf{Y},\mathbf{U})$ and $q(\boldsymbol{\theta}|\ q_1(\boldsymbol{\theta}_1)\ q_2(\boldsymbol{\theta}_2)\ldots q_K(\boldsymbol{\theta}_K)$, $D(q||g)$, is minimized [20], where $D(q||g)$ is defined as

$$D(q||g)=\int q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})}{g(\boldsymbol{\theta}|\nu,\mathbf{Y},\mathbf{U})}d\boldsymbol{\theta}$$

It can be shown that $q_k(\boldsymbol{\theta}_k)$ is expressed as

$$q_k(\boldsymbol{\theta}_k)=C\exp\{E_{-\boldsymbol{\theta}k}[\log g(\boldsymbol{\theta}|\nu,\mathbf{Y},\mathbf{U})]\} \tag{5}$$

where C is a normalized constant and $\mathrm{E}_{-\boldsymbol{\theta}k}[.]$ indicates the expectations of parameters other than $\boldsymbol{\theta}_k$ that are calculated with respect to every other parameter's variational posteriors except $q_k(\boldsymbol{\theta}_k)$ [21].

In the varBayes method that applies the variational approximation procedure to the multi-trait Bayesian model considered here, $g(\boldsymbol{\theta}|\nu,\mathbf{Y},\mathbf{U})$ is assumed to be approximated by a factorized function $q(\boldsymbol{\theta})$ that is written as

$$q(\boldsymbol{\theta})=q(\mathbf{b})q(\mathbf{\Sigma}_e)\prod_{l=1}^{N}\left\{q\left(\gamma_l,\mathbf{g}_l\right)q\left(\mathbf{\Sigma}_{gl}\right)\right\}q(\mathbf{S})$$

where we denote all of the variational posteriors for the different parameters in the right-hand side by $q(.)$ for simplicity with the understanding that $q(\mathbf{b})$, q $(\mathbf{\Sigma}_e)$ and so on take different forms depending on the parameters. The forms of these variational posteriors are derived from (5) in a manner similar to that used for derivation of the full conditional posteriors for the parameters in Gibbs sampling and the computational details are given in Additional file 2. In the following, we will give the variational posteriors for parameters, γ_l, $\mathbf{g}_l$, $\mathbf{\Sigma}_{gl}$ $(l=1,2,\ldots,N)$, $\mathbf{b}$, $\mathbf{\Sigma}_e$ and $\mathbf{S}$.

The variational posterior for $\mathbf{b}$, $q(\mathbf{b})$,is a multivariate normal density with mean

$$E(\mathbf{b}) = \{\sum_{i=1}^{n} \mathbf{X}_i' E(\boldsymbol{\Sigma}_e^{-1}) \mathbf{X}_i\}^{-1} \sum_{i=1}^{n} \mathbf{X}_i' E(\boldsymbol{\Sigma}_e^{-1}) \{\mathbf{y}_i - \sum_{l=1}^{N} u_{il} E(\gamma_l \mathbf{g}_l)\} \quad (6)$$

and variance-covariance matrix

$$V(\mathbf{b}) = \{\sum_{i=1}^{n} \mathbf{X}_i' E(\boldsymbol{\Sigma}_e^{-1}) \mathbf{X}_i\}^{-1}$$

where $E(.)$ indicates the expectation calculated with respect to the variational posteriors of relevant parameters, while $q(\boldsymbol{\Sigma}_e)$ is $\mathrm{IW}_T(n\text{-}T\text{-}1, \mathbf{S}_e)$ with

$$\mathbf{S}_e = \sum_{i=1}^{n} \{\mathbf{y}_i - \mathbf{X}_i E(\mathbf{b}) - \sum_{l=1}^{N} u_{il} E(\gamma_l \mathbf{g}_l)\} \{\mathbf{y}_i - \mathbf{X}_i E(\mathbf{b}) - \sum_{l=1}^{N} u_{il} E(\gamma_l \mathbf{g}_l)\}' \quad (7)$$

from which we obtain

$$E(\boldsymbol{\Sigma}_e^{-1}) = (n - T - 1)\mathbf{S}_e^{-1} \quad (8)$$

The variational posterior of $\boldsymbol{\Sigma}_{gl}$ is represented by $\mathrm{IW}_T(\nu_{gl}, \mathbf{S}_{gl})$, where

$$\nu_{gl} = \nu + E(\gamma_l)$$

and

$$\mathbf{S}_{gl} = E(\mathbf{S}) + E(\gamma_l \mathbf{g}_l \mathbf{g}_l')$$

The expectation of $\boldsymbol{\Sigma}_{gl}^{-1}$ with respect to this posterior distribution is

$$E(\boldsymbol{\Sigma}_{gl}^{-1}) = \nu_{gl} \mathbf{S}_{gl}^{-1} = \{\nu + E(\gamma_l)\}\{E(\mathbf{S}) + E(\gamma_l \mathbf{g}_l \mathbf{g}_l')\} \quad (9)$$

For γ_l and $\mathbf{g}_l$, we consider joint distribution for variational posterior $q(\gamma_l, \mathbf{g}_l)$ that is expressed as

$$q(\gamma_l, \mathbf{g}_l) \propto [(1-\pi)|\mathbf{V}_{gl}|^{1/2} |E(\boldsymbol{\Sigma}_l^{-1})|^{1/2} \exp\{\frac{1}{2} \hat{\mathbf{g}}_l' \mathbf{V}_{gl}^{-1} \hat{\mathbf{g}}_l\} \times \phi(\mathbf{g}_l | \hat{\mathbf{g}}_l, \mathbf{V}_{gl})]^{\gamma_l} \times [\pi \delta(\mathbf{0})]^{1-\gamma_l} \quad (10)$$

where $\phi(\mathbf{g}_l | \hat{\mathbf{g}}_l, \mathbf{V}_{gl})$ is a density function of a multivariate normal distribution $N(\hat{\mathbf{g}}_l, \mathbf{V}_{gl})$ with mean

$$\hat{\mathbf{g}}_l = \{E(\boldsymbol{\Sigma}_l^{-1}) + \sum_{i=1}^{n} u_{il}^2 E(\boldsymbol{\Sigma}_e^{-1})\}^{-1} E(\boldsymbol{\Sigma}_e^{-1}) \sum_{i=1}^{n} u_{il} \{\mathbf{y}_i - \mathbf{X}_i E(\mathbf{b}) - \sum_{j \neq l} u_{ij} E(\gamma_j \mathbf{g}_j)\} \quad (11)$$

and variance-covariance matrix

$$\mathbf{V}_{gl} = \{E(\boldsymbol{\Sigma}_l^{-1}) + \sum_{i=1}^{n} u_{il}^2 E(\boldsymbol{\Sigma}_e^{-1})\}^{-1} \quad (12)$$

The marginal posterior distribution of γ_l is a binomial distribution with probabilities of γ_l =1 and 0 given by

$$q(\gamma_l = 1) = \mathrm{E}(\gamma_l) = \frac{(1-\pi)|\mathbf{V}_{gl}|^{1/2} |\mathrm{E}(\boldsymbol{\Sigma}_l^{-1})|^{1/2} \exp\left\{\frac{1}{2} \hat{\mathbf{g}}_l' \mathbf{V}_{gl}^{-1} \hat{\mathbf{g}}_l\right\}}{(1-\pi)\left|\mathbf{V}_{gl}\right|^{1/2} |\mathrm{E}(\boldsymbol{\Sigma}_l^{-1})|^{1/2} \exp\left\{\frac{1}{2} \hat{\mathbf{g}}_l' \mathbf{V}_{gl}^{-1} \hat{\mathbf{g}}_l\right\} + \pi} \quad (13)$$

and

$$q(\gamma_l = 0) = 1 - q(\gamma_l = 1) = \frac{\pi}{(1-\pi)|\mathbf{V}_{gl}|^{1/2} |E(\boldsymbol{\Sigma}_l^{-1})|^{1/2} \exp\left\{\frac{1}{2} \hat{\mathbf{g}}_l' \mathbf{V}_{gl}^{-1} \hat{\mathbf{g}}_l\right\} + \pi}$$

The conditional distribution of $\mathbf{g}_l$ given γ_l is given by

$$q(\mathbf{g}_l | \gamma_l) = \begin{cases} N(\hat{\mathbf{g}}_l, \mathbf{V}_{gl}) & (\gamma_l = 1) \\ \delta(\mathbf{0}) & (\gamma_l = 0) \end{cases}$$

Therefore, we obtain

$$E(\gamma_l \mathbf{g}_l) = q(\gamma_l = 1) E(\mathbf{g}_l | \gamma_l = 1) + q(\gamma_l = 0) E(\mathbf{g}_l | \gamma_l = 0) = E(\gamma_l) \hat{\mathbf{g}}_l \quad (14)$$

and

$$E(\gamma_l \mathbf{g}_l \mathbf{g}_l') = E(\gamma_l)(\hat{\mathbf{g}}_l \hat{\mathbf{g}}_l' + \mathbf{V}_{gl}) \quad (15)$$

From (4), the variational posterior of $\mathbf{S}$ is a T-variate Wishart distribution with a scale matrix $\boldsymbol{\Sigma}_S$ and degree of freedom ν_s, $\mathrm{W}_T(\nu_s, \boldsymbol{\Sigma}_S)$, where $\boldsymbol{\Sigma}_s = \{\sum_{l=1}^{N} E(\boldsymbol{\Sigma}_{gl}^{-1})\}^{-1}$ and $\nu_s = N\nu + T + 1$, and the expectation of $\mathbf{S}$ is expressed as

$$E(\mathbf{S}) = \nu_s \boldsymbol{\Sigma}_S = (N\nu + T + 1)\{\sum_{l=1}^{N} E(\boldsymbol{\Sigma}_{gl}^{-1})\} - 1 \quad (16)$$

As outlined above, a well-known probability distribution, such as normal, inverse Wishart and so on, is assigned to the variational posterior of each parameter, which is characterized by the expectations of the functions of other parameters, taken with respect to their variational posteriors. The relationships between these expectations are given by (6), (8), (9), (13), (14), (15) and (16), from which the expectations can be calculated with numerical iterations to obtain the variational posteriors.

Moreover, the prior probability for a SNP to have zero effect, π, can be treated as a variable parameter to be inferred from the data as in BayesCπ and BayesDπ when

$0 < \pi < 1$. Here, we assume a uniform prior on $0 < \pi < 1$ to obtain a Beta distribution, B(a,b), as a variational posterior from (4), where $a = N - \sum_{l=1}^{N} E(\gamma_l) + 1$ and $b = \sum_{l=1}^{N} E(\gamma_l) + 1$, with the expectation

$$E(\pi) = a/(a+b) = \left\{N - \sum_{l=1}^{N} E(\gamma_l) + 1\right\}/(N+2) \quad (17)$$

Accordingly, the variational posteriors of the other parameters are modified with π substituted by $E(\pi)$ when π is involved in the Bayesian inference.

Treatment of missing phenotypes

In the phenotypic records for multiple traits, it is common for trait values to be partially missing in some individuals. Missing phenotypes of a trait in individuals can be inferred with the observed phenotypes of other traits in the same individuals. The step for inferring missing phenotypes can be implemented in the MCBayes and varBayes procedures as described below.

When there are missing phenotypes in an individual, the residual vector of the individual, **e** in model (1), is partitioned into components $\mathbf{e}_o$ and $\mathbf{e}_m$ corresponding to observed and missing phenotypes, respectively. Following [22], $\mathbf{e}_m$ can be sampled from the following normal distribution,

$$\mathbf{e}_m \sim N\left(\boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}{}^{-1}\mathbf{e}_o, \boldsymbol{\Sigma}_{mm} - \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}{}^{-1}\boldsymbol{\Sigma}_{mo}'\right) \quad (18)$$

where $\boldsymbol{\Sigma}_{mm}$, $\boldsymbol{\Sigma}_{mo}$ and $\boldsymbol{\Sigma}_{oo}$ indicate the partition of the residual variance-covariance matrix $\boldsymbol{\Sigma}_e$ corresponding to $\mathbf{e}_m$ and $\mathbf{e}_o$. Accordingly, $\mathbf{e}_m$ is drawn with Gibbs sampling in MCBayes while it is obtained as $E(\boldsymbol{\Sigma}_{mo})E(\boldsymbol{\Sigma}_{oo})^{-1}\mathbf{e}_o{}^*$ in varBayes, where, denoting the component corresponding to the missing traits by subscript 'o', $\mathbf{e}_o{}^*$ is written as

$$\mathbf{e}_o{}^* = \mathbf{y}_o - \mathbf{X}_o E(\mathbf{b}) - \sum_{l=1}^{N} u_{il} E\left(\gamma_l \mathbf{g}_{ol}\right)$$

and the expectations can be calculated with the variational posteriors of $\boldsymbol{\Sigma}_e$, **b** and $\mathbf{g}_l$. Missing phenotypes are inferred as the sum of the estimates of **b**, $\mathbf{g}_l$ and $\mathbf{e}_m$, which are used for the construction of the prediction model.

Simulation experiments

We simulated datasets to evaluate the accuracy of the predicted GBVs using the proposed Bayesian methods, MCBayes and varBayes, for multiple traits. In generating the datasets, three traits, denoted as A, B and C, were considered. The simulation of population and genome was carried out following [3] where a single trait was treated, while multiple traits were generated through a modified approach. The simulated population, genotypes and phenotypes are described in the following.

Populations with an effective population size of 100 were maintained by random mating for 1000 generations to attain mutation drift balance and LD between SNPs and QTLs. In generation 1001 and 1002, the population size was increased to 1000. The population in the 1001st generation was treated as a training population, where the phenotypes of three traits and SNP genotypes of the individuals were simulated and analyzed to estimate the SNP effects in the model. The phenotype of each trait for each individual in the 1001st generation was given as the sum of QTL effects over the polymorphic QTLs and environmental effects, which were sampled as described later. For simplicity, no other fixed effects were assumed. The population in the 1002nd generation was used as a test population, where the individuals were only genotyped for SNP markers without phenotypic records and GBVs of three traits were predicted for each individual using a model with SNP effects estimated based on the training population in the 1001st generation. The true breeding value (TBV) of the individual in the 1002nd generation was also simulated as the sum of QTL effects corresponding to the QTL genotype for each trait and used for evaluating the accuracy of predicted GBV, but was regarded as unknown and unavailable in the estimation of SNP effects in the models. Accuracy was measured based on the correlation between the TBV and predicted GBV, $r_{TBV,pGBV}$, for each trait and regression of TBV on predicted GBV, $b_{TBV,pGBV}$, was also obtained for assessing the bias of the prediction.

The genome was assumed to consist of 10 chromosomes each 100cM in length. Two scenarios were considered for the number of available SNP markers and the datasets under these two scenarios were denoted as Data I and Data II. In Data I, 101 marker loci were located every 1cM on each chromosome for a total of 1010 markers on a genome. In Data II, 1010 equidistant marker loci were located on each chromosome for a total of 10100 markers. We assumed that 100 equidistant QTLs were located on each chromosome such that a QTL was in the middle between two marker loci in both Data I and Data II. Therefore, there were a total of 1000 QTLs located on a whole genome. The mutation rates assumed per locus per meiosis were 2.5×10^{-3} and 5.0×10^{-5} for the marker locus and QTL, respectively. At least one mutation occurred in the most of the marker loci with a high mutation rate during the simulated generations. In the marker loci experiencing more than one mutation, the mutation remaining at the highest minor allele frequency (MAF) was regarded as visible, whereas the others were ignored, which resulted in the marker loci having two alleles similar to SNP markers. Although the mutation rate for QTL was assumed 2.5×10^{-5} in the simulation for a single trait conducted in [3], we here doubled it for generating TBVs of multiple traits for the reason as described below.

The polymorphic QTLs at which mutation occurred were used to simulate the three traits, A, B and C, the heritabilities of which, denoted by h_A^2, h_B^2 and h_C^2, respectively, were assumed to be $h_A^2 = 0.8$, $h_B^2 = 0.1$ and $h_C^2 = 0.1$. We divided all polymorphic QTLs into four groups, Group1, Group2, Group3 and Group4, according to the causal relationship with traits, where Group1 was a group of QTLs affecting both traits A and B, Group2 was a group of QTLs affecting only trait A, Group3 was a group of QTLs affecting only trait B and Group4 was a group of QTLs affecting only trait C. Therefore, it was assumed that QTLs of Group1 had pleiotropic effects on traits A and B while QTLs of other groups affected only a single trait. Each polymorphic QTL was randomly assigned to one of these groups, that is, to Group1, Group2, Group3 and Group4 with respective probabilities 0.4, 0.1, 0.1 and 0.4. Therefore, 80% of the QTL loci affecting trait A were shared by trait B on average and whereas all the QTL loci affecting trait C influence neither traits A nor B. In this setting of multi-trait case, the mutation rate of QTL was increased to 5.0×10^{-5} from 2.5×10^{-5}, the mutation rate adopted in [3] for generating a single trait, to retain the number of QTL per trait.

The pleiotropic effects of each QTL in Group1 were assumed to be correlated between traits A and B with correlation coefficient of 0.9. Consequently, genetic correlations between traits A and B, between A and C and between B and C, denoted as ρ_{GAB}, ρ_{GAC} and ρ_{GBC}, respectively, were $\rho_{GAB} = 0.72$ and $\rho_{GAC} = \rho_{GBC} = 0$ on average although the values of correlation coefficients were somewhat fluctuated in each data generation. This setting of traits was adopted to investigate how the prediction accuracy was increased for a low-heritability trait (trait B) by simultaneous analysis for multiple traits including a correlated high-heritability trait (trait A).

The effects of QTL alleles were sampled from gamma distributions independently for each QTL. Pleiotropic effects of QTL alleles in Group1 were determined for traits A and B by generating two correlated gamma random variables, x and y, with the correlation coefficient of 0.9, and assigning them with positive or negative values with equal probabilities. The effects of QTL alleles in the other groups, which affected each single trait only, was determined by sampling a gamma random variable z and by similarly assigning it with the positive or negative sign. We generated these three random variables x, y and z such that their marginal distributions were the same gamma distribution with scale parameter 0.4 and shape parameter 1.66, Gamma(0.4,1.66). For obtaining correlated variables x and y, we generated three independent gamma variables, x_1, x_2 and x_3, which were sampled from Gamma (0.36,1.66), Gamma(0.04,1.66) and Gamma(0.04,1.66), respectively, and determined the values of x and y as $x = x_1 + x_2$ and $y = x_1 + x_3$. It can be shown that x and y had a correlation coefficient of 0.9 and the same marginal distribution Gamma(0.4,1.66) [23].

The environmental correlation coefficients between three traits were denoted by ρ_{EAB}, ρ_{EAC} and ρ_{EBC}, respectively, and assumed as $\rho_{EAB} = 0.1$, $\rho_{EAC} = 0.2$ and $\rho_{EBC} = 0.3$. The environmental effects were sampled from a trivariate normal distribution with a mean vector $\mathbf{0}$ and a variance-covariance matrix $\mathbf{R}_E$, where

$$\mathbf{R}_E = \begin{pmatrix} \sigma_{EA}{}^2 & \rho_{EAB}\sigma_{EA}\sigma_{EB} & \rho_{EAC}\sigma_{EA}\sigma_{EC} \\ \rho_{EAB}\sigma_{EA}\sigma_{EB} & \sigma_{EB}{}^2 & \rho_{EBC}\sigma_{EB}\sigma_{EC} \\ \rho_{EAC}\sigma_{EA}\sigma_{EC} & \rho_{EBC}\sigma_{EB}\sigma_{EC} & \sigma_{EC}{}^2 \end{pmatrix}$$

with σ_{EA}^2, σ_{EB}^2 and σ_{EC}^2 indicating environmental variances of three traits. Environmental variance of trait A was given by $\sigma_{EA}^2 = (1/h_A^2\text{-}1)\sigma_{GA}^2$ with its heritability h_A^2 and genetic variance σ_{GA}^2, which was variance of TBV of trait A, and those of traits B and C were obtained similarly. The environmental effects were added to TBVs which were given by the sum of QTL effects to determine phenotypic values of three traits for individuals in the 1001st generation.

In Data I, 100 replicated datasets were simulated while Data II consisted of 20 replicated datasets due to the larger number of SNPs. Each of replicated datasets included records of phenotypes of three traits and genotypes of SNPs for the training population (1001st generation) and only SNP genotypes for the test population (1002nd generation). To simulate the situation of missing phenotypes, we generated additional datasets by deleting the phenotypic records of some traits for some individuals in the 100 replicated training datasets of Data I. These 100 replicated datasets were referred to as Data III, where the phenotypic records of traits A, B and C were respectively deleted for individuals of $i = 801\text{–}1000$, individuals of $i = 1\text{-}500$ and individuals of $i = 201\text{-}700$ in 1000 individuals ($i = 1\text{-}1000$) of the 1001st generation of Data I. Therefore, in Data III, the prediction model for GBVs was constructed with a training dataset consisting of the phenotypes of 800, 500 and 500 individuals for traits A, B and C, respectively, where only 100 individuals ($i = 701\text{-}800$) had phenotypic records of all three traits, and 1000 individuals in the 1002nd generation were predicted for GBV based on the genotypes of 1010 markers. The setting of non-zero environmental correlations, i.e., $\rho_{EAB} = 0.1$, $\rho_{EAC} = 0.2$ and $\rho_{EBC} = 0.3$, was adopted here to assess the benefit from the estimation process of missing phenotypes implemented in multi-trait analyses for the prediction accuracy, where the information of environmental covariance between observed and missing phenotypes was utilized as in (18).

Each replicated dataset in Data I, Data II and Data III was analyzed using the proposed methods for multiple traits, MCBayes and varBayes, to construct the GBV prediction model in the1001st generation and investigate

$r_{\mathrm{TBV,pGBV}}$ and $b_{\mathrm{TBV,\ pGBV}}$ for each trait in the 1002nd generation. For a comparison with conventional single-trait GBV prediction, the same datasets were also analyzed with the single-trait setting of MCBayes where three traits were separately treated without considering the correlation structure between traits.

We conducted cross-validation as well to evaluate the prediction performance within a population in the 1001st generation without the 1002nd generation since the techniques of cross-validation have commonly been used for evaluation of the accuracy in the studies of genomic selection with the actual datasets of animals [24-26] and plants [27-29], where the prediction accuracy of the model for the unobserved future samples was of concern. We applied 10-fold cross-validation. In brief, we randomly split a population of 1000 individuals in the 1001st generation into ten subpopulations with each size 100. By using a single subpopulation used as a test set and the remaining nine subpopulations as a training set to construct prediction model, the prediction accuracy and bias were assessed for the test set. This process was repeated ten times until all single subpopulations were used as test sets exactly once. We averaged the prediction accuracy and bias over ten repetitions to evaluate the prediction performance in each dataset. Because of much computational burden with MCMC analysis for repeated model constructions in cross-validation, evaluation with 10-fold cross-validation was carried out for Data I only, where mean of the prediction accuracies and biases in 100 datasets were obtained as $r_{\mathrm{TBV,pGBV}}$ and $b_{\mathrm{TBV,\ pGBV}}$ for each trait and for each prediction method. In the process of cross-validation, we also investigated the correlation between predicted GBV and the phenotypic value, $r_{\mathrm{y,pGBV}}$, in place of TBV, and regression of predicted GBV on phenotypic value, $b_{\mathrm{y,\ pGBV}}$, for each trait.

Two settings for the prior probability that a SNP has zero effect, π, were adopted in model construction, i.e., π was fixed at 0 or π was variable over the range $0 < \pi < 1$ and inferred from the data. The values of hyperparameter ν, degree of freedom of prior inverse Wishart distribution of $\mathbf{\Sigma}_{gl}$, were set to 5.0 and 3.2 corresponding to $\pi = 0$ and $0 < \pi < 1$ after preliminary analyses to evaluate the effects of the values of ν on prediction accuracy over $3 < \nu < 6$ although the accuracy was little affected by the values of ν.

In the MCMC iteration of MCBayes, we repeated 11000 cycles including a burn-in period of the first 1000 cycles. The values of parameters were sampled every 10 cycles to obtain the posterior means that were used to determine a prediction model for each generated dataset. In the method of varBayes, we adopted the criterion $|\hat{\boldsymbol{\theta}}^{*} - \hat{\boldsymbol{\theta}}|^2 / |\hat{\boldsymbol{\theta}}^{*}|^2 < 10^{-8}$ for convergence in numerical iteration for computing the expectations of parameters, where $\hat{\boldsymbol{\theta}}^{*}$ and $\hat{\boldsymbol{\theta}}$ are the current and previous value of the expectations of parameters. When this criterion was satisfied, the computational procedure with variational approximation was regarded as converged.

Results

We evaluated the accuracy and bias of the predicted GBVs, $r_{\mathrm{TBV,pGBV}}$ and $b_{\mathrm{TBV,\ pGBV}}$, obtained by the proposed methods for genomic selection of multiple traits including MCBayes and varBayes in 100 simulated datasets of Data I and Data III and in 20 datasets of Data II in comparison with the prediction accuracy with single-trait MCBayes analysis. The means of accuracies and biases of prediction evaluated using a test population in the 1002nd generation are summarized in Table 1, Table 2 and Table 3 for Data I, Data II and Data III, respectively. In all of datasets, MCBayes provided higher

Table 1 Accuracy and bias of predicted GBVs in Data I

Method			Trait A	Trait B	Trait C
MCBayes	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.788 ± 0.051	0.581 ± 0.103	0.453 ± 0.090
		$b_{\mathrm{TBV,pGBV}}$	0.994 ± 0.038	1.048 ± 0.264	1.00 ± 0.370
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.753 ± 0.060	0.580 ± 0.117	0.364 ± 0.137
		$b_{\mathrm{TBV,pGBV}}$	1.070 ± 0.064	1.149 ± 0.340	1.016 ± 0.364
varBayes	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.754 ± 0.061	0.570 ± 0.113	0.383 ± 0.117
		$b_{\mathrm{TBV,pGBV}}$	1.054 ± 0.051	0.994 ± 0.233	0.899 ± 0.247
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.716 ± 0.070	0.548 ± 0.122	0.347 ± 0.131
		$b_{\mathrm{TBV,pGBV}}$	0.894 ± 0.054	0.834 ± 0.186	0.636 ± 0.202
single-trait	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.783 ± 0.051	0.469 ± 0.083	0.455 ± 0.076
(MCBayes)		$b_{\mathrm{TBV,pGBV}}$	0.978 ± 0.037	1.020 ± 0.301	0.970 ± 0.259
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.778 ± 0.050	0.491 ± 0.114	0.483 ± 0.101
		$b_{\mathrm{TBV,pGBV}}$	1.089 ± 0.054	1.110 ± 0.634	1.061 ± 0.338

Averages and standard errors based on 100 replicates of simulated data are listed for prediction accuracy, $r_{\mathrm{pGBV,TBV}}$, and bias, $b_{\mathrm{pGBV,TBV}}$, of each trait. For the prior probability that a SNP has zero effect, $\boldsymbol{\pi}$, we considered two settings, in which $\boldsymbol{\pi}$ was fixed at 0, meaning the inclusion of all SNPs in the model, and $\boldsymbol{\pi}$ was varied over $0 < \boldsymbol{\pi} < 1$ and inferred from the data.

Table 2 Accuracy and bias of predicted GBVs in Data II

Method			Trait A	Trait B	Trait C
MCBayes	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.902 ± 0.032	0.706 ± 0.103	0.519 ± 0.097
		$b_{\mathrm{TBV,pGBV}}$	0.998 ± 0.034	0.902 ± 0.111	0.796 ± 0.179
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.868 ± 0.047	0.731 ± 0.120	0.401 ± 0.182
		$b_{\mathrm{TBV,pGBV}}$	1.092 ± 0.093	1.189 ± 0.199	1.198 ± 0.553
varBayes	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.859 ± 0.049	0.656 ± 0.110	0.438 ± 0.074
		$b_{\mathrm{TBV,pGBV}}$	1.059 ± 0.065	0.799 ± 0.105	0.724 ± 0.111
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.838 ± 0.061	0.678 ± 0.140	0.330 ± 0.157
		$b_{\mathrm{TBV,pGBV}}$	0.983 ± 0.034	0.851 ± 0.138	0.562 ± 0.155
single-trait	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.884 ± 0.039	0.485 ± 0.086	0.493 ± 0.089
(MCBayes)		$b_{\mathrm{TBV,pGBV}}$	0.974 ± 0.035	0.766 ± 0.113	0.766 ± 0.113
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.843 ± 0.044	0.597 ± 0.120	0.601 ± 0.109
		$b_{\mathrm{TBV,pGBV}}$	1.562 ± 0.261	1.787 ± 0.431	1.832 ± 0.565

Averages and standard errors based on 20 replicates of simulated data are listed for prediction accuracy, $r_{\mathrm{pGBV,TBV}}$, and bias, $b_{\mathrm{pGBV,TBV}}$, of each trait. For the settings of the prior probability that a SNP has zero effect, π, see Table 1.

prediction accuracy than varBayes in multi-trait analysis (Tables 1, 2 and 3). In Data I, $r_{\mathrm{TBV,pGBV}}$ obtained with multi-trait MCBayes analysis was 0.788 (0.051), 0.581 (0.103) and 0.453 (0.090) for traits A, B and C, respectively, when π was fixed at 0, where standard errors were given in parenthesis here and hereafter, while that was 0.753 (0.060), 0.580 (0.117) and 0.364 (0.137) when π was varied and inferred. In Data II with the number of SNPs increased to 10100, $r_{\mathrm{TBV,pGBV}}$ with multi-trait MCBayes analysis was higher at 0.902 (0.032), 0.706 (0.103) and 0.519 (0.097) for traits A, B and C, respectively, when π was fixed at 0 while that was 0.868 (0.047), 0.731 (0.120) and 0.401 (0.182) when π was inferred. With multi-trait varBayes analysis, $r_{\mathrm{TBV,pGBV}}$ was 0.754 (0.061), 0.570 (0.113) and 0.383 (0.117) for traits A, B and C with π =0 and that was 0.716 (0.070), 0.548 (0.122) and 0.347 (0.131) with π variable in Data I while that was 0.859 (0.049) , 0.656 (0.110) and 0.438 (0.074) with π =0 and 0.838 (0.061) , 0.678 (0.140) and 0.330 (0.157) with π variable in Data II. The prediction with multi-trait MCBayes was almost unbiased in Data I, where $b_{\mathrm{TBV,pGBV}}$ was near 1, but was biased in Data II for traits B and C (Tables 1 and 2). The analysis with varBayes showed greater bias than with MCBayes.

In single-trait analysis where MCBayes was applied for a single-trait model, $r_{\mathrm{TBV,pGBV}}$ was comparable with that of multi-trait analysis for high-heritability trait A while it was significantly decreased for low-heritability trait B which was highly correlated with trait A (Table 1 and Table 2). For trait C which had same heritability as trait B but was genetically independent of trait A ($\rho_{\mathrm{GAC}} = 0.0$), multi-trait MCBayes analysis gave the accuracy comparable to single-

Table 3 Accuracy and bias of predicted GBVs in Data III

Method			Trait A	Trait B	Trait C
MCBayes	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.766 ± 0.058	0.500 ± 0.127	0.322 ± 0.082
		$b_{\mathrm{TBV,pGBV}}$	0.977 ± 0.048	0.998 ± 0.356	0.967 ± 0.773
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.723 ± 0.069	0.503 ± 0.141	0.202 ± 0.119
		$b_{\mathrm{TBV,pGBV}}$	1.065 ± 0.076	1.195 ± 0.530	0.799 ± 0.523
varBayes	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.726 ± 0.072	0.447 ± 0.131	0.261 ± 0.134
		$b_{\mathrm{TBV,pGBV}}$	0.984 ± 0.052	0.582 ± 0.241	0.383 ± 0.181
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.679 ± 0.081	0.387 ± 0.115	0.228 ± 0.112
		$b_{\mathrm{TBV,pGBV}}$	0.840 ± 0.068	0.389 ± 0.132	0.240 ± 0.110
single-trait	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.760 ± 0.058	0.345 ± 0.070	0.336 ± 0.068
(MCBayes)		$b_{\mathrm{TBV,pGBV}}$	0.965 ± 0.047	0.931 ± 0.368	0.969 ± 0.510
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.758 ± 0.057	0.362 ± 0.105	0.354 ± 0.101
		$b_{\mathrm{TBV,pGBV}}$	1.086 ± 0.068	1.455 ± 1.401	1.251 ± 1.310

Averages and standard errors based on 100 replicates of simulated data are listed for prediction accuracy, $r_{\mathrm{pGBV,TBV}}$, and bias, $b_{\mathrm{pGBV,TBV}}$, of each trait. For the settings of prior probability that a SNP has zero effect, π, see Table 1.

trait MCBayes analysis when π =0 while, when π was varied and inferred, single-trait analysis showed considerably higher accuracy than multi-trait MCBayes and varBayes analyses as shown in Table 1 and Table 2. The prediction with single-trait analysis was almost unbiased In Data I, but that was likely to be biased in Data II with $b_{\mathrm{TBV,pGBV}}$ much deviated from 1 especially when π was varied and inferred.

In Data III with missing phenotypes, which was derived from Data I by removing some phenotypic records, $r_{\mathrm{TBV,\ pGBV}}$ was decreased for all traits in all methods due to the smaller sample size in comparison with Data I (Table 3). The rate of decrease in $r_{\mathrm{TBV,pGBV}}$ was greater for traits B and C than A as the greater reduction of the sample size was made in B and C. Multi-trait MCBayes analysis provided higher prediction accuracy than multi-trait varBayes analysis for all traits except for trait C withπ varied and inferred. For trait B, multi-trait analysis showed higher prediction accuracy than single-trait analysis, but gave less accurate prediction for trait C. For the bias of predicted GBV, $b_{\mathrm{TBV,\ pGBV}}$, the MCBayes analysis with π =0 was almost unbiased for both multi-trait and single-trait analyses, where $b_{\mathrm{TBV,\ pGBV}}$ ranged 0.931 to 0.998, although the standard error of $b_{\mathrm{TBV,\ pGBV}}$ was increased for low-heritability traits B and C, for which the MCBayes analyses with π varied and varBayes showed much biased prediction.

We listed $r_{\mathrm{TBV,\ pGBV}}$ and $b_{\mathrm{TBV,\ pGBV}}$ obtained by cross-validation conducted in a population of the 1001st generation of Data I as well as $r_{\mathrm{y,pGBV}}$ and $b_{\mathrm{y,pGBV}}$, correlation between predicted GBV and phenotype and regression of phenotype on predicted GBV, in Table 4. The prediction accuracy evaluated with cross-validation was considerably higher than that evaluated with a test population in the next generation, with the prediction bias being similarly for both evaluated with cross-validation and use of a test population. The relative merits in performance of prediction between the methods were also similar for both evaluations. It was shown in Table 4 that the relational expression, $r_{\mathrm{TBV,pGBV}} = r_{\mathrm{y,pGBV}}/h$ [30], held with h being a

Table 4 Accuracy and bias of predicted GBVs evaluated with cross-validation in Data I

Method			Trait A	Trait B	Trait C
MCBayes	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.832 ± 0.039	0.611 ± 0.095	0.501 ± 0.083
		$b_{\mathrm{TBV,pGBV}}$	1.016 ± 0.022	1.072 ± 0.231	1.052 ± 0.341
		$r_{\mathrm{y,pGBV}}$	0.741 ± 0.037	0.191 ± 0.045	0.160 ± 0.050
		$b_{\mathrm{y,pGBV}}$	1.013 ± 0.015	1.062 ± 0.125	1.039 ± 0.130
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.791 ± 0.048	0.603 ± 0.112	0.390 ± 0.127
		$b_{\mathrm{TBV,pGBV}}$	1.132 ± 0.064	1.210 ± 0.303	1.180 ± 0.379
		$r_{\mathrm{y,pGBV}}$	0.705 ± 0.045	0.191 ± 0.046	0.121 ± 0.060
		$b_{\mathrm{y,pGBV}}$	1.131 ± 0.065	1.210 ± 0.170	1.119 ± 0.373
varBayes	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.813 ± 0.049	0.620 ± 0.108	0.470 ± 0.118
		$b_{\mathrm{TBV,pGBV}}$	1.080 ± 0.048	0.994 ± 0.157	0.963 ± 0.201
		$r_{\mathrm{y,pGBV}}$	0.722 ± 0.047	0.187 ± 0.056	0.143 ± 0.058
		$b_{\mathrm{y,pGBV}}$	1.072 ± 0.049	0.945 ± 0.195	0.931 ± 0.289
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.779 ± 0.059	0.593 ± 0.111	0.423 ± 0.123
		$b_{\mathrm{TBV,pGBV}}$	0.944 ± 0.040	0.816 ± 0.139	0.662 ± 0.153
		$r_{\mathrm{y,pGBV}}$	0.690 ± 0.056	0.180 ± 0.055	0.125 ± 0.062
		$b_{\mathrm{y,pGBV}}$	0.935 ± 0.039	0.787 ± 0.166	0.626 ± 0.255
single-trait	$\pi = 0$	$r_{\mathrm{TBV,pGBV}}$	0.826 ± 0.040	0.515 ± 0.073	0.505 ± 0.074
(MCBayes)		$b_{\mathrm{TBV,pGBV}}$	0.997 ± 0.023	1.073 ± 0.270	1.030 ± 0.303
		$r_{\mathrm{y,pGBV}}$	0.735 ± 0.039	0.159 ± 0.045	0.162 ± 0.044
		$b_{\mathrm{y,pGBV}}$	0.993 ± 0.012	1.071 ± 0.162	1.055 ± 0.222
	$0 < \pi < 1$	$r_{\mathrm{TBV,pGBV}}$	0.821 ± 0.039	0.531 ± 0.099	0.522 ± 0.094
		$b_{\mathrm{TBV,pGBV}}$	1.131 ± 0.046	1.265 ± 0.555	1.192 ± 0.405
		$r_{\mathrm{y,pGBV}}$	0.731 ± 0.037	0.164 ± 0.051	0.164 ± 0.048
		$b_{\mathrm{y,pGBV}}$	1.127 ± 0.043	1.249 ± 0.482	1.205 ± 0.457

Averages and standard errors evaluated with 10- fold cross-validation are listed based on 100 replicates of simulated data in Data I are listed for prediction accuracy, $r_{\mathrm{pGBV,TBV}}$, and bias, $b_{\mathrm{pGBV,TBV}}$, as well as correlation between phenotypic value and predicted GBV, $r_{\mathrm{y,pGBV}}$, and regression of phenotypic value on predicted GBV, $b_{\mathrm{y,pGBV}}$, of each trait. For the settings of prior probability that a SNP has zero effect, π, see Table 1.

square-root of heritability and that $b_{\mathrm{TBV,pGBV}}$ was almost the same as $b_{\mathrm{y,pGBV}}$.

Correlation coefficients between predicted GBVs of trait A, B and C were listed in Table 5 for multi-trait MCBayes and varBayes analyses and single-trait MCBayes analysis in all datasets as well as the correlation coefficients between simulated breeding values (TBVs) of three traits in the 1002nd generation. Although the correlation of simulated breeding values between traits A and B was expected to be 0.72, the actual correlations obtained were biased upwards: 0.755 (0.133) and 0.735 (0.166) in Data I (III) and Data II, respectively. Multi-trait analysis could better estimate these correlations compared to single-trait analysis as seen in the correlation between predicted GBVs between traits A and B. For trait C which was genetically independent of A and B, the correlations between predicted GBVs for traits A and C and traits B and C did not significantly deviate from zero (Table 5).

Discussion

In this study, we proposed Bayesian methods for simultaneously predicting GBVs for multiple traits, where two computational procedures were devised using MCMC iteration and variational approximation, referred to as MCBayes and varBayes, respectively. A Bayesian model for simultaneously analyzing multiple traits was obtained by extending a Bayesian model for single-trait genomic selection proposed by [2] and [4] to multi-trait case. We introduced the prior probability that a SNP has zero effect, π, and accordingly, MCBayes with π fixed at 0, meaning that all SNPs are included in a model as covariates, constructs a model for multi-trait GBV prediction in a similar manner to BayesD. On the other hand, the MCMC procedure of MCBayes with π variable and inferred from the data is not exactly the same as that of BayesDπ, where a prior distribution of the variance and covariance matrix of SNP effects, $\mathbf{\Sigma}_{gl}$, is assumed to be independent of whether the SNP is included ($\gamma_l = 1$) or excluded($\gamma_l = 0$) from the model in MCBayes, as seen in (4), while it is dependent onγ_l and takes different forms for $\gamma_l = 1$ and $\gamma_l = 0$ in BayesDπ. This modification allows MCMC iteration of MCBayes to be performed with only Gibbs sampling.

In the simultaneous analysis of multiple traits for constructing a GBV prediction model, the computational burden greatly increases depending on the number of analyzed traits in comparison with single-trait analysis. We developed a variational approximation procedure, varBayes, for MCBayes to reduce the computational time for multi-trait analysis. In varBayes, the joint

Table 5 Correlations between predicted GBVs for trait A, B and C

Data	Method	A-B (0.72)	A-C (0.0)	B-C (0.0)
I	MCBayes $\pi = 0$	0.588 ± 0.181	−0.071 ± 0.175	−0.091 ± 0.199
	$0 < \pi < 1$	0.699 ± 0.188	−0.057 ± 0.250	−0.076 ± 0.301
	varBayes $\pi = 0$	0.688 ± 0.184	−0.100 ± 0.241	−0.077 ± 0.279
	$0 < \pi < 1$	0.644 ± 0.169	−0.053 ± 0.192	−0.058 ± 0.247
	Single-trait (MCBayes) $\pi = 0$	0.446 ± 0.132	0.058 ± 0.096	0.096 ± 0.121
	$0 < \pi < 1$	0.452 ± 0.183	0.035 ± 0.090	0.060 ± 0.111
	Simulated BV	0.755 ± 0.133	0.003 ± 0.054	0.004 ± 0.060
II	MCBayes $\pi = 0$	0.606 ± 0.197	−0.129 ± 0.113	−0.071 ± 0.140
	$0 < \pi < 1$	0.721 ± 0.222	−0.161 ± 0.149	−0.128 ± 0.179
	varBayes $\pi = 0$	0.614 ± 0.206	−0.176 ± 0.135	−0.090 ± 0.179
	$0 < \pi < 1$	0.671 ± 0.210	−0.153 ± 0.168	−0.047 ± 0.193
	Single-trait (MCBayes) $\pi = 0$	0.394 ± 0.115	0.020 ± 0.075	0.072 ± 0.111
	$0 < \pi < 1$	0.479 ± 0.206	−0.018 ± 0.084	0.022 ± 0.094
	Simulated BV	0.735 ± 0.166	−0.031 ± 0.054	−0.012 ± 0.041
III	MCBayes $\pi = 0$	0.530 ± 0.195	−0.032 ± 0.223	−0.037 ± 0.238
	$0 < \pi < 1$	0.662 ± 0.208	−0.014 ± 0.326	−0.013 ± 0.369
	varBayes $\pi = 0$	0.555 ± 0.194	−0.028 ± 0.218	−0.023 ± 0.248
	$0 < \pi < 1$	0.455 ± 0.167	−0.001 ± 0.158	0.021 ± 0.190
	Single-trait (MCBayes) $\pi = 0$	0.315 ± 0.106	0.029 ± 0.105	0.080 ± 0.134
	$0 < \pi < 1$	0.323 ± 0.138	0.024 ± 0.097	0.057 ± 0.130

Averages and standard errors are listed based on 100 replicates of simulated data in Data I and Data III and 20 replicates in Data II. Simulated BV indicates simulated breeding values, where expected correlations are 0.72, 0.0 and 0.0 for trait-pairs A-B, A-C and B-C as listed in parentheses. In Data III, correlations between simulated breeding values are the same as those in Data I.

posterior distribution of parameters was approximated by a factorized function, each component of which approximated marginal posterior distribution of each parameter and was referred to as a variational posterior. Variational posteriors were shown to be well-known distribution functions such as normal or inverse Wishart that could be derived by simple non-MCMC based numerical iteration.

In genomic selection, it is important to construct a model that enables accurate prediction for GVBs. Therefore, precise point estimation of the model parameters is more relevant rather than the construction of their posterior distributions. Accordingly, the evaluation of loss of prediction accuracy with varBayes in comparison with MCBayes would be suitable for the evaluation of approximation accuracy of varBayes. Using simulation experiments, we investigated the performance of the prediction model constructed with multi-trait analysis compared with single-trait analysis as well as the model constructed using variational approximation. Moreover, the performance of multi-trait analysis in the case of missing phenotypic records commonly occurring in the treatment of the actual data of multiple traits were evaluated based on the results of simulations. These points are discussed below including the computational time and the possible extension of prediction model considering polygenic effects.

Increase in accuracy for GBV prediction with multi-trait analysis

We evaluated the increase in prediction accuracy with multi-trait analysis in comparison with single-trait analysis using the datasets without missing phenotypes, Data I and Data II (Table 1 and Table 2). For trait A having heritability 0.8, multi-trait analysis and single-trait analysis made no difference in the prediction accuracy with MCBayes. Therefore, the advantage of multi-trait analysis over single-trait analysis in predicting GBV is negligible for high-heritability traits. However, we anticipate the increase in accuracy with multi-trait analysis for low-heritability traits utilizing correlations with high-heritability traits. Actually, for low-heritability trait B with heritability 0.1, which has highly correlated with trait A ($\rho_{GAB} = 0.72$), prediction accuracy was increased with multi-trait analysis. As trait C was not genetically correlated with trait A, the accuracy of predicting the GBV of C was not improved with multi-trait MCBayes analysis. The accuracy of predicted GBV of trait B was also increased with multi-trait varBayes analysis in Data I and Data II in comparison with single-trait MCBayes analysis while the prediction accuracy of trait C was lower with multi-trait varBayes analysis (Table 1 and Table 2). The genetic correlations between traits A and B were better estimated with predicted GBVs in multi-trait analysis than in single-trait analysis as shown in Table 5. Therefore, it can be concluded that low-heritability traits (heritability around 0.1) are better predicted for GBVs utilizing their correlations with high-heritability traits (heritability around 0.8), if any, with multi-trait analysis. However, the benefit of multi-trait analysis would be subtle for high-heritability traits. A similar finding was also reported in [13]. For uncorrelated low-heritability traits such as trait C, it is likely that the prediction using multi-trait analysis with π varied and inferred is less effective in comparison with single-trait analysis. Therefore, multi-trait analysis with π fixed at 0, which allows highly accurate prediction for correlated traits while retaining prediction accuracy comparable with single-trait analysis for uncorrelated low-heritability traits, would be a suitable method of choice for multi-trait genomic selection.

Approximation accuracy of variational procedure for MCMC estimation

Generally, constructing a GBV prediction model with MCMC estimation based on genotypic records for tens of thousands of SNPs and phenotypes for hundreds of individuals requires considerable computational time even for single-trait cases. Much more computational burden would be imposed in constructing a model in multi-trait analysis, depending on the number of traits of interest. Therefore, we proposed a computationally cost-effective method, varBayes, approximating MCMC based method, MCBayes, using a variational approximation procedure.

Simulation experiments showed that the prediction accuracy was lower with varBayes than with MCBayes in multi-trait analysis but the rate of loss of accuracy was not remarkable and was less than 10 percent for traits A and B under the same setting of π while it was greater for C (Table 1 and Table 2). The prediction accuracy for trait B correlated with high-heritability trait A was still higher with multi-trait varBayes analysis than with single-trait MCBayes analysis indicating that varBayes could well utilize the information on the correlation structure in multiple traits. Actually, multi-trait varBayes analysis could better capture the genetic correlation between A and B than single-trait MCbayes analysis (Table 5).

The computational time was greatly reduced for multi-trait varBayes analysis in comparison with multi-trait MCBayes analysis. We carried out all computations using a Fortran program written to implement multiple-trait analysis on a computer having two CPUs each with a quad-core processor (Intel Xeon 2.4GHz). In Data I, where 100 replicates of datasets each including genotypes of 1010 SNPs for 1000 individuals were simulated, varBayes took only 12 minutes with π fixed and 22 minutes with π varied for 100 times of model constructions while the computational time for MCBayes was respectively 440 and 435 minutes. Therefore, the average

computational time required to construct a prediction model for each dataset in Data I was less than 15 seconds for varBayes and was more than 4 minutes for MCBayes. For a larger data, Data II, that included 20 replicates of datasets each consisting of genotypes of 10100 SNPs for 1000 individuals, the computational time was 21 and 19 minutes for varBayes with π fixed and π varied, respectively, and the time for MCBayes were increased to more than 12 hours with the average computational times for each model construction being about 1 minute for varBayes and more than 30 minutes for MCBayes. In cross-validation process in which a total of one thousand repetitions of model constructions were performed in Data I, the computational time was about 3days with MCBayes while varBayes completed the analysis within only 4 hours.

Taking computational time and prediction accuracy into account, varBayes is considered a useful method for multi-trait genomic selection, which can rapidly construct a prediction model that is less accurate than that with the MCMC-based method for multi-trait analysis, but is more accurate than that with single-trait analysis for correlated traits. The usefulness of varBayes would be more remarkable for simultaneous prediction of GBVs of a large number of traits based on a huge number of SNPs where the application of an MCMC-based method might be prohibited.

Multi-trait analysis of dataset with missing phenotypes

In Data III, we simulated the datasets under the same condition as Data I except that some phenotypes were assumed to be unobserved. In short, we assumed that phenotypes of traits A, B and C were not available for 200, 500 and 500 individuals, respectively, in a total of 1000 individuals with only 100 individuals having the phenotypes of all three traits. In multi-trait analysis, missing phenotypes of individuals can be estimated with their observed phenotypes of other traits using (18), which indicates that residual effects of missing phenotypes can be restored from those of observed phenotypes. When the model fitting is successful for observed phenotypes, the residual effects of the phenotypes are well estimated by subtracting SNP effects and other fixed effects from the phenotypic effects, and those of missing phenotypes are suitably obtained by (18) utilizing the environmental correlation (covariance) between observed and missing phenotypes. Therefore, by assuming non-zero environmental correlation between traits ($\rho_{\mathrm{EAB}} = 0.1$, $\rho_{\mathrm{EAC}} = 0.2$, $\rho_{\mathrm{EBC}} = 0.3$), the loss of prediction accuracy caused by missing phenotypes was anticipated to be less with multi-trait analysis than with single-trait analysis. We expected that, in Data III, the prediction accuracy of trait C, environmentally correlated with trait A ($\rho_{\mathrm{EAC}} = 0.2$), was maintained higher in the presence of missing records for some phenotypes using multi-trait analysis in comparison with single-trait analysis. However, the rate of loss for the prediction accuracy for trait C was similar between multi-trait and single-trait analyses as seen in the comparison of the prediction accuracies between Data I and Data III (Tables 1 and 3). The utility of implementation of (18) for imputing missing phenotypes in the process of multi-trait model construction remains unclear in the settings of simulation adopted here.

Model extension by including polygenic effects

We can modify the Bayesian model (1) by including polygenic effects as follows;

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \sum_{l=1}^{N} \gamma_l u_{il}\mathbf{g}_l + \mathbf{v}_i + \mathbf{e}_i, \tag{19}$$

where $\mathbf{v}_i$ is a vector of polygenic effects for multiple traits and assumed to follow a multivariate normal distribution, $\mathbf{v}_i \sim N(\mathbf{0},\ \mathbf{\Sigma}_v)$ with $\mathbf{\Sigma}_v$ being a variance-covariance matrix of polygenic effects. The polygenic effects for all individuals of a training population and a test population are collectively denoted by $\mathbf{V}$, then the variance-covariance matrix of $\mathbf{V}$ can be expressed as $\mathbf{A} \otimes \mathbf{\Sigma}_v$, where $\mathbf{A}$ is additive genetic relationship matrix for analyzed individuals, computed from the information of pedigree or markers. When the low-density markers are used for the analysis, a considerable portion of genetic effects could not be captured by markers. Accordingly, if pedigree information is available, inclusion of polygenic effects estimated based on pedigree is beneficial in predicting breeding values for genomic selection. The revised model (19) can be similarly treated by MCBayes and varBayes as the model (1), where estimation steps for additional covariates and parameters, $\mathbf{V}$ and $\mathbf{\Sigma}_v$, can be easily implemented in the procedures of Bayesian multi-trait model construction using MCMC iteration and variational approximation. When no pedigree information is available, $\mathbf{A}$ is computed from marker information as a genomic relationship matrix. However, the model (19) includes all available markers as covariates, resulting in redundantly using the same marker information in the model fitting. In the availability of high-density SNP markers, which is the case for some species of animals and plants currently, the genetic relationships between individuals can be well captured by markers themselves in the multi-locus model (1), the benefit from the inclusion of the polygenic effects in the model would be subtle [31].

Conclusion

In this study, we described a statistical model for Bayesian simultaneous prediction of GBVs in genomic selection

targeting multiple traits and devised an MCMC-based method and a computationally cost-effective method utilizing the variational approximation procedure, referred to as MCBayes and varBayes, respectively, to estimate parameters included in the model. The results of simulation experiments showed that the multi-trait analysis that could utilize the correlation structure between traits allowed more accurate prediction of GBVs for correlated traits compared to single-trait analysis that treated each trait separately, where, for low-heritability traits correlated with high-heritability traits, the prediction accuracy for GBVs was remarkably improved with multi-trait analysis. Although the prediction accuracy with varBayes was lower than that with MCBayes in multi-trait analysis, the rate of loss in accuracy was moderate and the accuracy for correlated low-heritability traits was still higher with varBayes analysis compared to single-trait analysis. Considering the benefit of greatly reduced computational time, varBayes was considered to be a practical method for predicting GBVs in multi-trait genomic selection.

Additional files

Additional file 1: Derivation of full conditional posterior distributions of parameters in a statistical model.

Additional file 2: Derivation of variational posteriors for parameters in a statistical model.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TH devised Bayesian prediction methods for the genomic selection of multiple traits, developed a program for simulations and drafted the manuscript. HI assisted in developing a program and drafted the final manuscript. Both authors read and approved the final manuscript.

Acknowledgements

This study was supported by the Grant-in-Aid for Scientific Research (B) of Japan Society for the Promotion of Science (Grant No. 22380010).

Author details

[1]Agroinformatics Division, National Agriculture and Food Research Organization, Agricultural Research Center, Kannondai, Tsukuba, Ibaraki 305-8666, Japan. [2]Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo, Tokyo 113-8657, Japan.

Received: 22 August 2012 Accepted: 22 January 2013
Published: 31 January 2013

References

1. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES: **A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.** *PLoS One* 2011, **6**(5):e19379.
2. Meuwissen THE, Hayes B, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819–1829.
3. Solberg TR, Sonesson AK, Wooliams JA, Meuwissen THE: **Genomic selection using different marker types and densities.** *J Anim Sci* 2008, **86:**2447–2454.
4. Habier D, Fernando RL, Kizilkaya K, Garrick DJ: **Extension of the Bayesian alphabet for genomic selection.** *BMC Bioinformatics* 2011, **12:**186.
5. Xu S: **Estimating polygenic effects using makers of the entire genome.** *Genetics* 1989, **163:**789–891.
6. George EI, McCullogh RE: **Variable selection via Gibbs sampling.** *J Am Stat Assoc* 1993, **88:**881–889.
7. Gianola D, Delos Campos G, HillW G, Manfredi E, Fernando R: **Additive genetic variability and the Bayesian alphabet.** *Genetics* 2009, **183:**347–363.
8. Xu C, Wang X, Li Z, Xu S: **Mapping QTL for multiple traits using Bayesian statistics.** *Genet Res* 2008, **90:**1–15.
9. Banerjee S, Yandell BS, Yi N: **Bayesian Quantitative Trait Loci Mapping for Multiple Traits.** *Genetics* 2008, **179:**2275–2289.
10. Meuwissen THE, Goddard ME: **Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data.** *Genet Sel Evol* 2004, **36:**261–279.
11. Carlin BP, Chib S: **Bayesian model choice via Markov chain Monte Carlo.** *J R Stat Soc Ser B* 1995, **57:**473–484.
12. Yi N: **A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci.** *Genetics* 2004, **167:**967–975.
13. Calus MP, Veerkamp RF: **Accuracy of multi-trait genomic selection using different methods.** *Genet Sel Evol* 2011, **43:**26.
14. Yi N, Banerjee S: **Hierarchical generalized linear models for multiple quantitative trait locus mapping.** *Genetics* 2009, **181:**1101–1113.
15. Logsdon BA, Hoffman G, Mezey J: **A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis.** *BMC Bioinformatics* 2010, **11:**58.
16. Li Z, Sillanpää MJ: **Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms.** *Genetics* 2012, **190:**231–249.
17. Hayashi T, Iwata H: **EM algorithm for Bayesian estimation of genomic breeding values.** *BMC Genet* 2010, **11:**3.
18. Shepherd RK, Meuwissen THE, Wooliams JA: **Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers.** *BMC Bioinformatics* 2010, **11:**529.
19. Attias H: *Inferring parameters and structure of latent variable models by variational Bayes*, Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan-Kaufmann; 1999:21–30.
20. Bishop CM: *Pattern recognition and machine learning*. New York: Springer–Verlag; 2006.
21. Beal M: *Variational algorithms for approximate Bayesian inference*. University of London: PhD thesis; 2003.
22. VanTassel CP, VanVleck LD: **Multiple-trait Gibbs sampler for animal models: Flexible programs for Bayesian and likelihood-based (co) variance component inference.** *J Anim Sci* 1996, **74:**2586–2597.
23. Mathai AM, Moschopoulos PG: **A form of multivariate gamma distribution.** *Ann Inst Stat Math* 1992, **44:**97–106.
24. Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW: **A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers.** *Genet Sel Evol* 2009, **41:**56.
25. Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen THE: **The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation.** *Genetics* 2009, **183:**1119–1126.
26. Saatchi M, McClure MC, McKay SD, Megan M, Rolf MM, Kim JW, Decker JE, Taxis TM, Chapple RH, Ramey HR, *et al*: **Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation.** *Genet Sel Evol* 2011, **43:**40.
27. Lorenzana RE, Bernardo R: **Accuracy of genotypic value predictions for marker-based selection in biparental plant populations.** *Theor Appl Genet* 2009, **120:**151–161.
28. Crossa J, Delos Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, *et al*: **Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers.** *Genetics* 2010, **186:**713–724.
29. Heffner EL, Jannink JL, Iwata H, Souza E, Sorrells ME: **Genomic selection accuracy for grain quality traits in biparental wheat populations.** *Crop Sci* 2011, **51:**2597–2606.
30. Dekkers JCM: **Prediction of response to marker-assisted and genomic selection using selection index theory.** *J Anim Breed Genet* 2007, **124:**331–341.
31. Kärkkäinen HP, Sillanpää MJ: **Robustness of Bayesian multilocus association models to cryptic relatedness.** *Hum Genet* 2012, **76:**510–523.

doi:10.1186/1471-2105-14-34
Cite this article as: Hayashi and Iwata: **A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits.** *BMC Bioinformatics* 2013 **14**:34.