GSE Genetics Selection Evolution

RESEARCH Open Access

A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses

Rohan L Fernando[1*], Jack CM Dekkers[1] and Dorian J Garrick[1,2]

Abstract

Background: To obtain predictions that are not biased by selection, the conditional mean of the breeding values must be computed given the data that were used for selection. When single nucleotide polymorphism (SNP) effects have a normal distribution, it can be argued that single-step best linear unbiased prediction (SS-BLUP) yields a conditional mean of the breeding values. Obtaining SS-BLUP, however, requires computing the inverse of the dense matrix **G** of genomic relationships, which will become infeasible as the number of genotyped animals increases. Also, computing **G** requires the frequencies of SNP alleles in the founders, which are not available in most situations. Furthermore, SS-BLUP is expected to perform poorly relative to variable selection models such as BayesB and BayesC as marker densities increase.

Methods: A strategy is presented for Bayesian regression models (SSBR) that combines all available data from genotyped and non-genotyped animals, as in SS-BLUP, but accommodates a wider class of models. Our strategy uses imputed marker covariates for animals that are not genotyped, together with an appropriate residual genetic effect to accommodate deviations between true and imputed genotypes. Under normality, one formulation of SSBR yields results identical to SS-BLUP, but does not require computing **G** or its inverse and provides richer inferences. At present, Bayesian regression analyses are used with a few thousand genotyped individuals. However, when SSBR is applied to all animals in a breeding program, there will be a 100 to 200-fold increase in the number of animals and an associated 100 to 200-fold increase in computing time. Parallel computing strategies can be used to reduce computing time. In one such strategy, a 58-fold speedup was achieved using 120 cores.

Discussion: In SSBR and SS-BLUP, phenotype, genotype and pedigree information are combined in a single-step. Unlike SS-BLUP, SSBR is not limited to normally distributed marker effects; it can be used when marker effects have a *t* distribution, as in BayesA, or mixture distributions, as in BayesB or BayesCπ. Furthermore, it has the advantage that matrix inversion is not required. We have investigated parallel computing to speedup SSBR analyses so they can be used for routine applications.

Background

Due to advances in molecular biology, high-density single nucleotide polymorphisms (SNP) data are now being incorporated with phenotypic data into genetic evaluation [1-4] in what has been called genomic prediction or genomic selection [5]. Typically, genotypes are initially available only on a few thousand individuals at many thousands to several hundred thousand SNPs. Phenotypic values or deregressed estimated breeding values (EBV) on these genotyped individuals are used to estimate the effects of the SNPs using Bayesian multiple-regression models in which the marker effects are treated as random [5]. We refer to such models as marker effect models (MEM). The estimated marker effects are then used to predict the breeding values (BV) of animals that may not yet have phenotypes but have been genotyped.

Nejati-Javaremi [6] proposed an alternative approach to incorporate genotype information into genetic evaluation, where the BV of the animals are fitted, as in a pedigree-based best linear unbiased prediction (BLUP) analysis, but

*Correspondence: rohan@iastate.edu
[1]Department of Animal Science, Iowa State University, 50011 Ames, Iowa, USA
Full list of author information is available at the end of the article

with a genomic relationship matrix, computed from available genotypes, that replaces the pedigree-based additive relationship matrix. We refer to these models as breeding value models (BVM). Fernando [7] showed that the MEM and BVM provide equivalent predictions of BV and that predictions could be more easily obtained from MEM than from BVM because at the time of their study the number of markers was much smaller than the number of genotyped animals. This situation changed when high-density SNP data became available and the BVM was rediscovered for genomic selection [8,9]. In this context, BLUP using the BVM is often referred to as G-BLUP.

When MEM are used and not all animals are genotyped, marker-based EBV are typically combined with pedigree-based BLUP EBV from the entire breeding population to improve accuracy, using various selection-index approximations [2,10]. Legarra et al. [11-14] proposed an alternative to this approximate approach to combine information from genotyped and non-genotyped animals, where in a single step, they obtain BLUP EBV combining phenotypic, pedigree and SNP data using Henderson's mixed model equations (MME) for a BVM with a modified version of the additive relationship matrix **H** that reflects the additional information from the SNP genotypes. Thus, their method is called single-step BLUP (SS-BLUP), and is expected to yield unbiased predictions under multivariate normality, even in populations that are undergoing selection and non-random mating. This is important, because genotypes are usually collected only on superior animals and this can lead to biased evaluations. A properly calculated BLUP evaluation has been shown to be free of this selection bias [15-21]. Because their BLUP analysis is based on a BVM, we will refer to their SS-BLUP as SSBV-BLUP.

In SSBV-BLUP, the SNP data are used to construct the matrix **G** of genomic relationships for the genotyped individuals [6,8,22]. Conceptually, the remaining relationship coefficients constructed from pedigree are modified to provide consistency with **G** [11]. Provided that $\mathbf{G}^{-1}$ and the inverse $\mathbf{A}_{\mathbf{22}}^{-\mathbf{1}}$ of the corresponding additive relationship matrix are available, an efficient algorithm has been developed to construct $\mathbf{H}^{-1}$ [13]. However, computing $\mathbf{G}^{-1}$ and $\mathbf{A}_{\mathbf{22}}^{-\mathbf{1}}$ is an inefficient process, because the computing time to obtain these inverses is proportional to n_2^3, where n_2 is the number of individuals with genotypes. Nevertheless, when n_2 is a few thousands, SSBV-BLUP provides an elegant and convenient method to estimate BV that combines the available phenotype, pedigree and SNP data. Due to wide-spread adoption of genotyping in livestock, however, n_2 is becoming too large for SSBV-BLUP to remain computationally feasible much longer [23,24]. One strategy to overcome this problem is to approximate **G** such that the inverse could be computed efficiently [24]. Another strategy is to obtain solutions to the MME without explicitly inverting **G** [23]. Because **G** is very dense and it grows in size as more individuals are genotyped, these strategies are not very promising.

As discussed by Stranden and Garrick [9], when the number of genotyped individuals exceeds the number of marker covariates, use of MEM, which do not require computing **G** or its inverse, will lead to more efficient calculations. At present, however, most analyses that use MEM are based on Markov chain Monte-Carlo (MCMC) techniques that are computationally demanding. In addition, analyses using MEM have not been able to accommodate animals without genotypes.

In this paper, we extend MEM to accommodate animals without genotypes and propose alternative MCMC approaches to address computing requirements. The methodology presented here will extend the attractive features of SSBV-BLUP to Bayesian multiple-regression analyses that draw inferences from posterior distributions using MCMC techniques. Our extended MEM will enable BLUP evaluations without having to compute **G** or its inverse, while combining information from genotyped and non-genotyped animals.

Methods

We begin this section with a short introduction to the most widely used MEM and equivalent BVM. Then, we briefly review the theory that underlies SSBV-BLUP and show that the BVM used in SSBV-BLUP is equivalent to an MEM that can be used for single-step Bayesian regression (SSBR). Finally, strategies will be presented to implement a Gibbs sampler for drawing inferences from the posterior distributions of the breeding values.

Marker effect models

The groundbreaking paper of Meuwissen et al. [5] proposed three multiple-regression, MEM for genomic selection. These models can be described by the following general model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{T}\boldsymbol{\alpha} + \mathbf{e}, \tag{1}$$

where **y** is the vector of trait phenotypes, **X** is a known incidence matrix that relates the vector of non-genetic, "fixed" effects to **y**, $\mathbf{T} = \mathbf{M} - E(\mathbf{M})$, **M** is a matrix of marker covariates, $\boldsymbol{\alpha}$ is the vector of random, partial-regression coefficients of the marker covariates, and **e** is a vector of residuals. The expected value of the marker covariates can be written as $E(\mathbf{M}) = \mathbf{1}\mathbf{k}'$, where the row vector $\mathbf{k}'$ is the vector of expected values of marker covariates for a random animal in the absence of selection.

Meuwissen et al. [5] actually used haplotype covariates in their model, but now most analyses are based on SNP marker covariates. Their model assumed that the markers completely capture the first and second moments of the BV. When this is not true, a polygenic residual BV

can be added to equation (1). Strandén and Christensen showed that different allele coding methods lead to the same inference of marker effects when the general mean is included in the model and all genotypes are observed [25]. However, centered allele coding, as in matrix **T**, had a numerical advantage when MCMC methods were used [25].

In the Bayesian implementation of the MEM, the fixed effects are usually given a flat prior. The α_j are a priori assumed independently distributed as:

$$\alpha_j|\pi,\sigma^2_{\alpha_j} = \begin{cases} 0 & \text{with probability } \pi \\ \sim \mathrm{N}\left(0,\sigma^2_{\alpha_j}\right) & \text{with probability } (1-\pi), \end{cases} \tag{2}$$

where $\sigma^2_{\alpha_j}$ are a priori assumed identically and independently distributed (iid) scaled inverted chi-square variables with scale S^2_α and degrees of freedom ν_α. The residuals are typically assumed iid normal with null mean and variance σ^2_e, with a scaled inverted chi-square prior for σ^2_e, with scale S^2_e and degrees of freedom ν_e.

In the first model considered by Meuwissen et al. [5], the α_j were iid normal variables with null mean and a common "known" variance, which is equivalent to our general model with $\pi = 0$, $\nu_\alpha = \infty$, and S^2_α set to the common known variance. This model was called "BLUP" in their paper [5]. The second model that they considered (BayesA) is equivalent to our general model with $\pi = 0$, and S^2_α and ν_α set to "known" values. Their third model (BayesB) is identical to BayesA, except with $\pi = 0.95$ or some other "known" value.

Kizilkaya et al. [26,27] modified the "BLUP" model to have a value of $\pi > 0$ and an unknown common variance with a scaled inverted chi-square prior and referred to this model as BayesC. Furthermore, Habier et al. [27] extended the BayesC model in which the value of π is unknown with a uniform prior and referred to this as BayesCπ. In Bayesian Lasso, the marker effects are assigned a double exponential prior. This can be achieved by setting π in equation (2) to 0 and using an exponential prior distribution for $\sigma^2_{\alpha_j}$ [28].

Inferences on the breeding values and other unknown parameters in the model are made from their marginal posterior distributions, using MCMC methods [5,27,28]. Let $\mathbf{t}'_c$ be the row vector of marker covariates for some selection candidate. Then, the conditional mean for its genomic EBV is $\mathbf{t}'_c\hat{\boldsymbol{\alpha}}$, where $\hat{\boldsymbol{\alpha}}$ is the posterior mean of $\boldsymbol{\alpha}$, which can be computed from the MCMC samples.

Breeding value models

Two mixed linear models are said to be linearly equivalent if the vector **y** of observations has the same first and second moments in both models [29]. In that sense, a model that is equivalent to (1) can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \mathbf{e}, \tag{3}$$

where $\mathbf{g} = \mathbf{T}\boldsymbol{\alpha}$, and $\mathbf{T} = \mathbf{M} - E(\mathbf{M})$ is the matrix of centered marker covariates. One of the advantages of using centered marker covariates is that the vector **g** of breeding values will have null means even if $\boldsymbol{\alpha}$ does not have null means. The covariance matrix of the breeding values is:

$$\begin{aligned} \mathrm{Var}(\mathbf{g}|\mathbf{T}) &= \mathrm{Var}(\mathbf{T}\boldsymbol{\alpha}) \\ &= \mathbf{T}\mathrm{Var}(\boldsymbol{\alpha})\mathbf{T}'. \end{aligned}$$

Then, in both models (1) and (3), the mean of **y** is $\mathbf{X}\boldsymbol{\beta}$ and the covariance matrix is:

$$\mathrm{Var}(\mathbf{y}|\mathbf{T}) = \mathbf{T}\mathrm{Var}(\boldsymbol{\alpha})\mathbf{T}' + \mathbf{I}\sigma^2_e.$$

Thus, these two models are linearly equivalent and the parameters of one model can always be written as linear functions of the parameters of the other model. "Consequently, linear and quadratic estimates under one model can be converted by these same linear functions to estimates of an equivalent model" [29].

When the number of markers is large relative to the size of **g**, BLUP of **g** can be obtained efficiently [8,9] by solving the MME that correspond to model (3). When $\pi = 0$ and $\nu_\alpha = \infty$ in (1), the covariance matrix of marker effects is:

$$\mathrm{Var}(\boldsymbol{\alpha}) = \mathbf{I}\sigma^2_\alpha,$$

and the covariance matrix of the genomic BV conditional on **M** can be written as:

$$\mathrm{Var}(\mathbf{g}|\mathbf{T}) = \mathbf{T}\mathbf{T}'\sigma^2_\alpha. \tag{4}$$

Furthermore, under some assumptions, the variance σ^2_α of marker effects can be related to the variance σ^2_g of BV as:

$$\sigma^2_\alpha = \frac{\sigma^2_g}{\sum_j 2p_j(1-p_j)}, \tag{5}$$

where p_j is the frequency of SNP j [22,30,31]. Then, equation (4) can be written as:

$$\begin{aligned} \mathrm{Var}(\mathbf{g}|\mathbf{T}) &= \frac{\mathbf{T}\mathbf{T}'}{\sum_j 2p_j(1-p_j)}\sigma^2_g \\ &= \mathbf{G}\sigma^2_g. \end{aligned} \tag{6}$$

Suppose genotypes were not available and the analysis is conditional only on pedigree, which we denote as **P**. Then, the conditional mean of **g** given **P** is null and the conditional covariance matrix is:

$$\mathrm{Var}(\mathbf{g}|\mathbf{P}) = \mathbf{A}\sigma^2_g,$$

where **A** is the additive relationship matrix, and the variance is computed over the conditional distribution of **T**. Using this variance for **g** in setting up the MME for

equation (3) results in the usual non-genomic MME for the BVM.

Theory underlying SSBV-BLUP

Legarra et al. [11] proposed an ingenious strategy to combine information from genotyped and non-genotyped animals in a single BLUP analysis based on a BVM, which we refer to as SSBV-BLUP. Suppose $\mathbf{g}$ is partitioned as:

$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{T}_2\boldsymbol{\alpha} \end{bmatrix},$$

where $\mathbf{g_1}$ are BV of the animals with missing genotypes $\mathbf{T}_1$ and $\mathbf{g}_2$ are BV of those with observed genotypes $\mathbf{T}_2$. Following Legarra et al. [11], the vector $\mathbf{g}_1$ can be written as:

$$\begin{aligned} \mathbf{g}_1 &= \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2 + \left(\mathbf{g}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2\right) \\ &= \hat{\mathbf{g}}_1 + \boldsymbol{\epsilon}, \end{aligned} \tag{7}$$

where $\mathbf{A}_{ij}$ are partitions of $\mathbf{A}$ that correspond to $\mathbf{g}_1$ and $\mathbf{g}_2$. The first term in equation (7) is the best linear predictor (BLP) of $\mathbf{g}_1$ given $\mathbf{g}_2$, and the second is a residual genetic effect to accommodate deviations between the true breeding value, $\mathbf{g}_1$, and its prediction from $\mathbf{g}_2$, $\hat{\mathbf{g}}_1$, which we refer to as $\boldsymbol{\epsilon}$, the "imputation residual".

Consider first the conditional distribution of $\mathbf{g}_1$ given $\mathbf{P}$. Then, as expected, the variance of $\mathbf{g}_1$ is:

$$\begin{aligned} \mathrm{Var}(\mathbf{g_1}|\mathbf{P}) &= \left[\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)\right]\sigma_g^2 \quad (8) \\ &= \mathbf{A}_{11}\sigma_g^2, \quad (9) \end{aligned}$$

where the first term of equation (8) is the variance of the $\hat{\mathbf{g}}_1$ (i.e. predicted from its relatives in $\mathbf{g}_2$) and the second term is the variance of $\boldsymbol{\epsilon}$. Similarly, $\mathrm{Var}(\mathbf{g}_2|\mathbf{P}) = \mathbf{A}_{22}\sigma_g^2$.

In this situation, where the covariance structure of $\mathbf{g}_1$ and $\mathbf{g}_2$ is determined entirely by the pedigree, it is easy to see that $\boldsymbol{\epsilon}$ in equation (7) is uncorrelated to $\mathbf{g}_2$, and therefore if $\mathbf{g}_1$ and $\mathbf{g}_2$ are multivariate normal, $\boldsymbol{\epsilon}$ and $\mathbf{g}_2$ are independent. Multivariate normality of $\mathbf{g}_1$, $\mathbf{g}_2$ and consequently of $\boldsymbol{\epsilon}$ will be a good approximation if the effective number of loci that contribute to the BV is large. This will be the case even when the individual marker effects do not have a normal distribution, as in BayesA and BayesB.

Consider now the conditional distribution of $\mathbf{g}_1$ given $\mathbf{T}_2$. Note that, given the observed genotypes $\mathbf{T}_2$, the distribution of $\mathbf{g}_2$ changes to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{T}_2\mathbf{T}_2'\sigma_\alpha^2$. Now, to obtain the result in [11] for the conditional distribution of $\mathbf{g}_1$ given $\mathbf{T}_2$, we have to assume that the change in the distribution of $\mathbf{g}_1$ occurs entirely as a result of the change in the distribution of $\mathbf{g}_2$. In other words, we only use the information in $\mathbf{T}_2$ that flows to $\mathbf{g}_1$ through $\mathbf{g}_2$ (see Discussion). Then, it can be reasoned that:

$$\begin{aligned} \mathrm{Var}\left(\mathbf{g}_1|\mathbf{T}_2\right) = {} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{T}_2\mathbf{T}_2'\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\sigma_\alpha^2 \\ & + \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)\sigma_g^2, \end{aligned} \tag{10}$$

where now the vector $\hat{\mathbf{g}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2$ has covariance matrix given by the first term of equation (10), and because $\boldsymbol{\epsilon}$ is independent of $\mathbf{g}_2$, the second term of equation (10) remains identical to that of equation (8). Similarly, the covariance between $\mathbf{g}_1$ and $\mathbf{g}_2$ conditional on $\mathbf{T}_2$ is:

$$\mathrm{Cov}(\mathbf{g}_1, \mathbf{g}_2|\mathbf{T}_2) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{T}_2\mathbf{T}_2'\sigma_\alpha^2.$$

Furthermore, assuming equation (5), the above results can be combined to show that conditional on $\mathbf{T}_2$, $\mathbf{g}$ has a multivariate normal distribution with null mean and covariance matrix:

$$\begin{aligned} &\mathrm{Var}(\mathbf{g}|\mathbf{T}_2) = \mathbf{H} \\ &= \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right) & \mathbf{A_{12}}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathrm{A}_{21} & \mathbf{G} \end{bmatrix}\sigma_g^2, \end{aligned} \tag{11}$$

where $\mathbf{G} = \mathbf{T}_2\mathbf{T}_2'/[\sum 2p_i(1-p_i)]$. An alternative derivation of this matrix was given by Christensen and Lund [14]. The inverse of this matrix is needed to set up the MME, and can be computed as

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A_{22}^{-1}} \end{bmatrix},$$

which is the basis for SSBV-BLUP [11,13,14]. Note that this requires that both $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$ are computed beforehand, which is not computationally feasible because these matrices are dense and large when the number n_2 of genotyped animals is large. Furthermore, $\mathbf{T}_2\mathbf{T}_2'$ is not full rank when n_2 exceeds the number of markers. Thus to obtain a full rank $\mathbf{G}$, ad-hoc adjustments are often made, such as adding small values to the diagonals or regressing $\mathbf{G}$ towards $\mathbf{A}$, which is justified when the markers do not capture all the genetic variability. However, due to the increased adoption of SNP genotyping in livestock, $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$ are becoming too large for SSBV-BLUP to remain computationally feasible [23,24]. A second problem in SSBV-BLUP is related to the scaling that is done using the SNP frequencies. As mentioned earlier, when all data that were used for selection are available for computing the conditional mean, it can be computed as if selection had not taken place [18,20,21]. If selection has taken place, this requires using SNP frequencies from the founders, because these frequencies are not changed by selection. However, in most situations SNP genotypes are not available in the founders and frequencies observed in

the genotyped animals are used, which can lead to biased evaluations, particularly in a multi-breed context.

Single-step Bayesian regression

Assuming that BV are captured completely by marker genotypes, the mixed linear model for the phenotypic values can be expressed in terms of a BVM (12) or an MEM (13) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e} \quad (12)$$
$$= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{T}\boldsymbol{\alpha} + \mathbf{e}, \quad (13)$$

where we have introduced the incidence matrix $\mathbf{Z}$ to accommodate animals with repeated records or animals without records. As in SSBV-BLUP, suppose $\mathbf{T}_1$ is not observed. Then it is not possible to use equation (13) as the basis for the MEM. Note that $\mathbf{T}_1\boldsymbol{\alpha}$ is equal to $\mathbf{g}_1$. So, using equation (7) for $\mathbf{g}_1$ and writing $\mathbf{g}_2 = \mathbf{T}_2\boldsymbol{\alpha}$, the model for the phenotypic values becomes:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\mathbf{Z}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{bmatrix} + \mathbf{e} \quad (14)$$
$$= \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\mathbf{Z}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{T}_2\boldsymbol{\alpha} + \boldsymbol{\epsilon}\\ \mathbf{T}_2\boldsymbol{\alpha}\end{bmatrix} + \mathbf{e} \quad (15)$$

To construct the matrix $\mathbf{T}_2$ of centered marker covariates, we need to know the value of $E(\mathbf{M}_2) = \mathbf{1}\mathbf{k}'$, where the row vector $\mathbf{k}'$ is the vector of expected values of marker covariates for a random animal in the absence of selection. However, marker covariates are often available only for animals that have been subject to selection. So, we propose to write $\mathbf{g}_2$ as:

$$\mathbf{g}_2 = \mathbf{T}_2\boldsymbol{\alpha} \quad (16)$$
$$= (\mathbf{M}_2 - \mathbf{1}\mathbf{k}')\boldsymbol{\alpha} \quad (17)$$
$$= \mathbf{M}_2\boldsymbol{\alpha} - \mathbf{1}\mu_g, \quad (18)$$

where $\mu_g = \mathbf{k}'\boldsymbol{\alpha}$ is assigned a flat prior. The reason μ_g can be considered an independent parameter is that $\mathbf{k}'$ is a vector of unknown parameters. Substituting equation (18) for $\mathbf{g}_2 = \mathbf{T}_2\boldsymbol{\alpha}$ in equation (15) gives:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1^*\\ \mathbf{X}_2^*\end{bmatrix}\boldsymbol{\beta}^* + \begin{bmatrix}\mathbf{Z}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\hat{\mathbf{M}}_1\boldsymbol{\alpha} + \boldsymbol{\epsilon}\\ \mathbf{M}_2\boldsymbol{\alpha}\end{bmatrix} + \mathbf{e} \quad (19)$$
$$= \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{W}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\epsilon} + \mathbf{e}, \quad (20)$$

where $\mathbf{X}_1^* = [\mathbf{X}_1, \mathbf{J}_1]$, $\mathbf{J}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{J}_2$, $\mathbf{X}_2^* = [\mathbf{X}_2, \mathbf{J}_2]$, $\mathbf{J}_2 = -\mathbf{1}$, $\hat{\mathbf{M}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2$,

$$\boldsymbol{\beta}^* = \begin{bmatrix}\boldsymbol{\beta}\\ \mu_g\end{bmatrix},\ \mathbf{U} = \begin{bmatrix}\mathbf{Z}_1\\ \mathbf{0}\end{bmatrix},$$
$$\mathbf{X}^* = \begin{bmatrix}\mathbf{X}_1^*\\ \mathbf{X}_2^*\end{bmatrix} \text{ and } \mathbf{W} = \begin{bmatrix}\mathbf{W}_1\\ \mathbf{W}_2\end{bmatrix} = \begin{bmatrix}\mathbf{Z}_1\hat{\mathbf{M}}_1\\ \mathbf{Z}_2\mathbf{M}_2\end{bmatrix}.$$

The matrix $\hat{\mathbf{M}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2$ of imputed marker covariates can be obtained efficiently, using partitioned inverse results, by solving the easily formed very sparse system [see Additional file 1]:

$$\mathbf{A}^{11}\hat{\mathbf{M}}_1 = -\mathbf{A}^{12}\mathbf{M}_2, \quad (21)$$

where $\mathbf{A}^{ij}$ are partitions of $\mathbf{A}^{-1}$ that correspond to partitioning $\mathbf{g}$ into $\mathbf{g}_1$ and $\mathbf{g}_2$. Similarly, the vector $\mathbf{J}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{J}_2$ can be obtained efficiently by solving the system:

$$\mathbf{A}^{11}\mathbf{J}_1 = -\mathbf{A}^{12}\mathbf{J}_2. \quad (22)$$

The differences between this MEM (20) and the model that is currently used for Bayesian regression (BR) are: (i) estimation of an extra fixed effect: $\mu_g = \mathbf{k}'\boldsymbol{\alpha}$, (ii) some of the marker covariates in (20) are imputed, and (iii) a residual term $\boldsymbol{\epsilon}$ has been introduced to account for deviations of the imputed marker covariates from their unobserved, actual values.

When genotypes are not missing, the vector $\mathbf{J}_1$ is null, and the covariate for μ_g only contains $\mathbf{J}_2$, which is in the column space of $\mathbf{X}_2$ when the model explicitly or implicitly contains the general mean. In this case, it has been shown that inference on differences between breeding values is not affected by the choice of vector $\mathbf{k}'$ used to center the marker covariates [25]. This includes using $\mathbf{k}' = \mathbf{0}'$, which corresponds to not centering, and therefore, μ_g does not have to be included in the model. However, when genotypes are missing $\mathbf{J}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{J}_2$ may not be in the column space of $\mathbf{X}_1$, and thus μ_g must be included in the model if $\mathbf{k}'$ is not known and the marker covariates are not centered. A thorough investigation of this approach of including μ_g as a fixed effect in the model in place of centering the marker covariates is beyond the scope of this paper, but a small simulation is included here to compare the accuracy of prediction when marker covariates are centered with those when marker covariates are not centered but μ_g is included in the model.

Regardless of the prior used for $\boldsymbol{\alpha}$, the distribution of the vector $\boldsymbol{\epsilon}$ of imputation residuals will be approximated by a multivariate normal vector with null mean and covariance matrix $(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})\sigma_g^2$ (see equation 10), where σ_g^2 is assigned a scaled inverse chi-square distribution with scale parameter S_g^2 and degrees of freedom ν_g. Imputing the marker covariates needs to be done only once, and it can be done efficiently in parallel. Imputation of unobserved marker covariates will not significantly increase the overall computing time, and the storage costs will not be greater than for centered observed genotypes.

The MME that correspond to equation (20) for BayesC with $\pi = 0$ are

$$\begin{bmatrix} \mathbf{X}^{*'}\mathbf{X}^* & \mathbf{X}^{*'}\mathbf{W} & \mathbf{X}_1^{*'}\mathbf{Z}_1 \\ \mathbf{W}'\mathbf{X}^* & \mathbf{W}'\mathbf{W} + \mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} & \mathbf{W}_1'\mathbf{Z}_1 \\ \mathbf{Z}_1'\mathbf{X}_1^* & \mathbf{Z}_1'\mathbf{W}_1 & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}^* \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'^*\mathbf{y} \\ \mathbf{W}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y}_1 \end{bmatrix}. \tag{23}$$

The submatrix of these MME that correspond to $\boldsymbol{\epsilon}$ are identical to those for $\mathbf{g}_1$ from a pedigree-based analysis and are very sparse. Thus, as explained in the next section, conditional on $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}$, $\boldsymbol{\epsilon}$ can be sampled efficiently by using either a blocking-Gibbs sampler [32,33] or a single-site Gibbs sampler, as used in pedigree-based analyses [33]. Note that these MME, which do not have $\mathbf{G}$ or its inverse, may be used to overcome the computational problems with SSBV-BLUP. The predicted BV can be written as:

$$\hat{\mathbf{g}} = \begin{bmatrix} \mathbf{J}_1 \\ \mathbf{J}_2 \end{bmatrix} \hat{\mu}_g + \begin{bmatrix} \hat{\mathbf{M}}_1 \\ \mathbf{M}_2 \end{bmatrix} \hat{\boldsymbol{\alpha}} + \mathbf{U}\hat{\boldsymbol{\epsilon}} \tag{24}$$

The MME given by equation (23) have the advantage that they will not grow in size as more animals are genotyped, in contrast to the MME corresponding to equation (14) that are given by Aguilar et al. [13]. Results from solving equation (23) will not be identical to those from the MME corresponding to equation (14) because in equation (23), $\mathbf{k}'$ is treated as an unknown, and $\mu_g = \mathbf{k}'\boldsymbol{\alpha}$ is estimated from the genotypic and phenotypic data. However, the MME corresponding to:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* \\ \mathbf{X}_2^* \end{bmatrix} \boldsymbol{\beta}^* + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} + \mathbf{e}, \tag{25}$$

where $\mathbf{a}_1 = \hat{\mathbf{M}}_1\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, and $\mathbf{a}_2 = \mathbf{M}_2\boldsymbol{\alpha}$ will give predictions for breeding values computed as:

$$\tilde{\mathbf{g}} = \begin{bmatrix} \mathbf{J}_1 \\ \mathbf{J}_2 \end{bmatrix} \hat{\mu}_g + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{a}}_1 \\ \hat{\mathbf{a}}_2 \end{bmatrix}, \tag{26}$$

identical to those from equation (24).

Numerical example

Consider the pedigree in Table 1, and assume genotypes are available only on individuals 1, 2, and 4. Genotypes $\mathbf{M}_2$ at 10 markers are in Table 2. Following Legarra et al. [11], the relationship matrix is rearranged such that $\mathbf{A}_{11}$ are relationships among individuals 3, 5, and 6, which do not have genotypes, and $\mathbf{A}_{22}$ are relationships among 1, 2, and 4, which have the genotypes in Table 2. The inverse of the rearranged relationship matrix is given in Table 3. The imputed genotypes, $\hat{\mathbf{M}}_1$, and the marker covariates, $\mathbf{J}_1$, for μ_g of the non-genotyped animals, could be obtained efficiently by solving the sparse systems (21) and (22), respectively (Table 4). To set up the MME, we will assume that $\sigma_\alpha^2 = \frac{\sigma_g^2}{10}$, $9\sigma_g^2 = \sigma_e^2$, and that μ, and μ_g are the only fixed effects. Then, the MME (23) and solutions corresponding to the MEM (20) are in Table 5. For comparison, the MME and solutions for the single-step BV model are given in Table 6. The solutions for μ and μ_g are identical for the two sets of MME. The BLUP of $\mathbf{g}$ obtained as equation (24), using the solutions to equation (23), are identical to those obtained from equation (26), and are in Table 1.

Table 1 Pedigree used in the numerical example

Individual	Sire	Dam	Phenotypes	SS-BLUP-BV
1	0	0	–	1.61
2	0	0	1.25	1.59
3	0	0	-0.34	0.00
4	1	2	1.30	1.62
5	1	2	1.27	1.61
6	1	3	0.46	0.80

Genotypes are available for individuals 1, 2 and 4. Phenotypes are available for all individuals except individual 1, the sire. Single-step, BLUP predictions of breeding values (SS-BLUP-BV) are in the last column.

Simulation to compare accuracy of prediction with and without centering of covariates

Accuracy of prediction was computed for SS-BLUP with and without centering of marker covariates. As demonstrated by the preceding numerical example, the MEM given by (20) or the BVM given by (25) can be used to get identical SS-BLUP predictions. When marker covariates were not centered, accuracy was computed with and without fitting μ_g in the model. The simulated data consisted of 20 paternal halfsib families, each with 20 offspring. Thus, the pedigree consisted of 20 unrelated sires, 400 unrelated dams and 400 offspring. Only genotypes from the 400 offspring were used in the analysis. Phenotypes from all 400 offspring and from 210 dams were used. Results are given for four scenarios of the simulation. In all four scenarios, the trait was determined by 50 QTL and had a heritability of 0.5. All QTL effects were sampled from a normal distribution with either mean $\mu_\alpha = 0$ or mean $\mu_\alpha = 0.2$. The analysis was based on either 100 or 10 000 marker genotypes, including the QTL. Correlations between the predicted and true breeding values for the three analyses and the four scenarios are in Table 7,

Table 2 Observed genotypes at ten markers for individuals in the example in Table 1

Individual	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10
1	1	2	1	1	0	0	1	2	1	0
2	2	1	1	1	2	0	1	1	1	1
4	1	1	0	1	1	0	2	1	2	1

Table 3 Inverse of rearranged relationship matrix for individuals in the example in Table 1

	3	5	6	1	2	4
3	1.50	0.00	-1.00	0.50	0.00	0.00
5	0.00	2.00	0.00	-1.00	-1.00	0.00
6	-1.00	0.00	2.00	-1.00	0.00	0.00
1	0.50	-1.00	-1.00	2.50	1.00	-1.00
2	0.00	-1.00	0.00	1.00	2.00	-1.00
4	0.00	0.00	0.00	-1.00	-1.00	2.00

Row and column labels are the individual identifiers.

which contains correlations for the non-genotyped animals, and Table 8, which contains the correlations for the genotyped animals. All three models had almost identical accuracies except when the number of observations exceeded the number of markers and the marker effects had a non-null mean. In this case, the model with centered marker covariates had a higher accuracy. The same accuracy could be achieved by including an extra covariate to model the mean of the breeding values. The results in this simulation indicate that this is necessary only when the number of observations exceeds the number of markers. When the number of markers greatly exceeds the number of observations, this extra covariate may become unnecessary even when the marker effects have a non-null mean.

Strategies to implement a Gibbs sampler

In most implementations of BR, a Gibbs sampler is used to draw inferences from the posterior distribution of the unknowns [5,27,33,34]. This involves sampling from full-conditional posterior distributions. Sampling of fixed effects, $\boldsymbol{\beta}$, is almost identical to sampling the marker effects, $\boldsymbol{\alpha}$ [34]. Thus, we will describe the strategy for sampling $\boldsymbol{\alpha}$ directly.

Sampling marker effects and variances

The most time-consuming task in a BR analysis is sampling $\boldsymbol{\alpha}$ from its full-conditional distribution. Detailed derivation of these full-conditionals and illustrative R scripts are in Fernando and Garrick [34] for BR models with complete genotype data. As shown in equation (10) of [34], the first step in sampling α_i is adjusting $\mathbf{y}$ for all other effects in the model. In the case of equation (20), $\tilde{\mathbf{y}}$, the vector of adjusted phenotypic values is:

$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^* - \sum_{j \neq i} \mathbf{w}_j \alpha_j - \mathbf{U}\boldsymbol{\epsilon}, \tag{27}$$

computed with the current values of $\boldsymbol{\beta}^*$, $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$. This vector is used to compute the right-hand-side for α_i, $rhs_i = \mathbf{w}_i'\tilde{y}$, which is needed to sample α_i in BayesA, BayesB and BayesC [5,34]. In BayesB and BayesC, α_i is either null or, conditional on the effect variance, normally distributed. Thus before α_i is sampled, a Bernoulli variable δ_i is sampled that indicates whether α_i is null or is normally distributed. Sampling δ_i requires computing the full-conditional probability that $\delta_i = 1$, which also requires rhs_i [34].

Calculation of rhs_i can be done more efficiently [35] by initially computing:

$$\hat{\mathbf{y}} = (\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^* - \mathbf{W}\boldsymbol{\alpha} - \mathbf{U}\boldsymbol{\epsilon}), \tag{28}$$

which is the vector $\mathbf{y}$ corrected for all effects in the model, using their current values. Then, before sampling α_i, the right-hand-side for α_i, $rhs_i = \mathbf{w}_i\tilde{y}$, is obtained efficiently as:

$$rhs_i = \mathbf{w}_i'\hat{\mathbf{y}} + (\mathbf{w}_i'\mathbf{w}_i)\alpha_i^{[\text{old}]}, \tag{29}$$

and after sampling α_i, $\hat{\mathbf{y}}$ is updated as:

$$\hat{\mathbf{y}} = \hat{\mathbf{y}} + \mathbf{w}_i' \left(\alpha_i^{[\text{old}]} - \alpha_i^{[\text{new}]}\right), \tag{30}$$

where $\alpha_i^{[\text{old}]}$ and $\alpha_i^{[\text{new}]}$ are the values of α_i before and after sampling. Then, sampling proceeds to the next locus. Thus for each marker covariate, rhs_i is always computed, whereas in BayesB and BayesC, the vector $\hat{\mathbf{y}}$ only needs to be updated when $\alpha_i^{[\text{old}]} \neq \alpha_i^{[\text{new}]}$. In computing rhs_i, the first term is the dot product of two vectors of length n, the number of records. In the second term, the scalar $\mathbf{w}_i'\mathbf{w}_i$ does not change from one round of sampling to the next, and so it is computed only once for each marker covariate. To speed-up computations, the dot product can be done in parallel, using the message passing interface (MPI) [36]. Suppose m processors are used for computing. Then, the first q elements of each covariate will be stored in memory of the first processor and the second set of q elements in the second processor and so on, where q is the whole part of $\frac{n}{m}$, and the last processor gets the remaining elements. Similarly, blocks of $\hat{\mathbf{y}}$ will also be stored with the m processors. Then, to compute $\mathbf{w}_i'\hat{\mathbf{y}}$, each of the $m-1$ processors will do a dot product of length q and the last processor one of length $\leq q$. Only the scalar result of the dot product from each processor needs to be communicated for the sampling and this will be relatively fast. After sampling α_i, updating $\hat{\mathbf{y}}$ will also be done in parallel. Before updating $\hat{\mathbf{y}}$, the scalar $\left(\alpha_i^{[\text{old}]} - \alpha_i^{[\text{new}]}\right)$ must be available to all m processors. Then, each processor will update its own

Table 4 Imputed genotypes at ten markers and covariates for μ_g for individuals in the example in Table 1

Individual	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	μ_g
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00
5	1.5	1.5	1.0	1.0	1.0	0.0	1.0	1.5	1.0	0.5	-1.00
6	0.5	1.0	0.5	0.5	0.0	0.0	0.5	1.0	0.5	0.0	-0.50

Table 5 Mixed model equations for marker effects model with observed and imputed marker covariates for the example in Table 1

	μ	μ_g	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	ϵ_3	ϵ_5	ϵ_6
μ	5.00	-3.50	5.00	4.50	2.50	3.50	4.00	0.00	4.50	4.50	4.50	2.50	1.00	1.00	1.00
μ_g	-3.50	3.25	-4.75	-4.00	-2.25	-3.25	-4.00	0.00	-4.25	-4.00	-4.25	-2.50	0.00	-1.00	-0.50
m_1	5.00	-4.75	8.61	5.75	3.75	4.75	6.50	0.00	5.75	5.75	5.75	3.75	0.00	1.50	0.50
m_2	4.50	-4.00	5.75	6.36	3.00	4.00	4.50	0.00	5.00	5.25	5.00	2.75	0.00	1.50	1.00
m_3	2.50	-2.25	3.75	3.00	3.36	2.25	3.00	0.00	2.25	3.00	2.25	1.50	0.00	1.00	0.50
m_4	3.50	-3.25	4.75	4.00	2.25	4.36	4.00	0.00	4.25	4.00	4.25	2.50	0.00	1.00	0.50
m_5	4.00	-4.00	6.50	4.50	3.00	4.00	7.11	0.00	5.00	4.50	5.00	3.50	0.00	1.00	0.00
m_6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
m_7	4.50	-4.25	5.75	5.00	2.25	4.25	5.00	0.00	7.36	5.00	6.25	3.50	0.00	1.00	0.50
m_8	4.50	-4.00	5.75	5.25	3.00	4.00	4.50	0.00	5.00	6.36	5.00	2.75	0.00	1.50	1.00
m_9	4.50	-4.25	5.75	5.00	2.25	4.25	5.00	0.00	6.25	5.00	7.36	3.50	0.00	1.00	0.50
m_{10}	2.50	-2.50	3.75	2.75	1.50	2.50	3.50	0.00	3.50	2.75	3.50	3.36	0.00	0.50	0.00
ϵ_3	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.17	0.00	-0.11
ϵ_5	1.00	-1.00	1.50	1.50	1.00	1.00	1.00	0.00	1.00	1.50	1.00	0.50	0.00	1.22	0.00
ϵ_6	1.00	-0.50	0.50	1.00	0.50	0.50	0.00	0.00	0.50	1.00	0.50	0.00	-0.11	0.00	1.22
rhs	3.94	-4.04	5.93	4.91	2.75	4.04	5.06	0.00	5.34	4.91	5.34	3.18	-0.34	1.27	0.46
sol	-0.34	-1.61	-0.01	-0.00	-0.01	-0.00	-0.01	0.00	0.01	-0.00	0.01	0.00	-0.00	0.00	-0.01

The last two rows give the right-hand-side and the solutions of the equations.

block of $\hat{\mathbf{y}}$ and will be ready for sampling the effect of the next marker covariate. The remaining calculations related to sampling marker effects do not depend on the number of records and will take only a negligible amount of time. Once the marker effects are sampled, sampling the locus-specific variances of marker effects in BayesA and BayesB, or the common variance in BayesC does not depend on the number of observations [5,34].

Parallel computations

To investigate the speedup from parallel computations, the Lonestar Linux cluster of the Texas Advanced Computing Center was used. According to the documentation at (http://www.tacc.utexas.edu/user-services/user-guides/lonestar-user-guide#overview) a regular compute node on this cluster contains two Xeon Intel Hexa-Core 64-bit Westmere processors (12 cores in all) on a single board, as an SMP unit. The core frequency is 3.33 GHz and supports four floating-point operations per clock period, with a peak performance of 13.3 GFLOPS/core or 160 GFLOPS/node. Each node contains 24 GB of memory (2 GB/core). The memory subsystem has three channels from each processor's memory controller to 3 DDR3 ECC DIMMS, running at 1333 MHz. The processor interconnect, QPI, runs at 6.4 GT/s.

Table 6 Mixed model equations for single-step BV model for the example in Table 1

	μ	μ_g	a_3	a_5	a_6	a_1	a_2	a_4
μ	5.00	-3.50	1.00	1.00	1.00	0.00	1.00	1.00
μ_g	-3.50	3.25	0.00	-1.00	-0.50	0.00	-1.00	-1.00
a_3	1.00	0.00	1.17	0.00	-0.11	0.06	0.00	0.00
a_5	1.00	-1.00	0.00	1.22	0.00	-0.11	-0.11	0.00
a_6	1.00	-0.50	-0.11	0.00	1.22	-0.11	0.00	0.00
a_1	0.00	0.00	0.06	-0.11	-0.11	0.32	-0.01	-0.09
a_2	1.00	-1.00	0.00	-0.11	0.00	-0.01	1.31	-0.17
a_4	1.00	-1.00	0.00	0.00	0.00	-0.09	-0.17	1.29
rhs	3.94	-4.04	-0.34	1.27	0.46	0.00	1.25	1.30
sol	-0.34	-1.61	-0.00	-0.01	-0.01	-0.00	-0.02	0.01

The last two rows give the right-hand-side and the solutions of the equations.

Table 7 Correlation between predicted and true breeding values of non-genotyped animals for three models

Markers	μ_α	Correlations		
		CC	CNμ_g	CN
100	0.0	0.67	0.67	0.66
100	0.2	0.67	0.67	0.59
10,000	0.0	0.60	0.60	0.60
10,000	0.2	0.58	0.58	0.58

Models with marker covariates centered (CC), marker covariates not centered with μ_g in the model (CNμ_g), and marker covariates not centered without μ_g (CN) in the model. The QTL effects were sampled from a normal distribution with either mean $\mu_\alpha = 0$ or mean $\mu_\alpha = 0.2$. The analyses were based on either 100 or 10 000 marker genotypes, including the QTL.

Table 8 Correlation between predicted and true breeding values of genotyped animals for three models

Markers	μ_α	Correlations		
		CC	CNμ_g	CN
100	0.0	0.93	0.93	0.91
100	0.2	0.92	0.92	0.78
10,000	0.0	0.71	0.71	0.71
10,000	0.2	0.76	0.76	0.76

Models with marker covariates centered (CC), marker covariates not centered with μ_g in the model (CNμ_g), and marker covariates not centered without μ_g (CN) in the model. The QTL effects were sampled from a normal distribution with either mean $\mu_\alpha = 0$ or mean $\mu_\alpha = 0.2$. The analyses were based on either 100 or 10 000 marker genotypes, including the QTL.

Initially, a data set with 1 million individuals and 5000 markers on each individual was used. Obtaining 100 samples of $\boldsymbol{\alpha}$ and σ^2_α for BayesC with $\pi = 0$ took 1167 seconds on a single core. By extrapolation, ignoring memory limitations, 100 samples for 50 000 markers would take about 11 670 seconds on a single core. Next, MPI [36] was used for parallel computation on 120 cores across 10 nodes. Then, obtaining 100 samples for 1 million individuals with 50 000 markers took 202 seconds. Thus, the speedup on 120 cores was about 58 times, and obtaining 40 000 samples would take about 22 hours.

Sampling imputation residuals and genetic variance

Using results from [33], it can be shown that conditional on the sampled value of all other variables and the data, the conditional distribution of $\boldsymbol{\epsilon}$ is multivariate normal with mean $\tilde{\boldsymbol{\epsilon}}$ that is given by the solution to the following system:

$$\left(\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2}\right)\tilde{\boldsymbol{\epsilon}} = \mathbf{Z}_1'(\mathbf{y}_1 - \mathbf{X}_1^*\boldsymbol{\beta}^* - \mathbf{W}_1\boldsymbol{\alpha}), \quad (31)$$

and covariance matrix: $\left(\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2}\right)^{-1}\sigma_e^2$. The right-hand-side of equation (31) can be obtained efficiently as $\mathbf{Z}_1'\hat{\mathbf{y}}_1 + \mathbf{Z}_1'\mathbf{Z}_1\boldsymbol{\epsilon}$, where $\hat{\mathbf{y}}_1$ is the current value of the sub-vector from equation (28) that corresponds to $\mathbf{y}_1$. A sample from this distribution can be obtained by using the blocking-Gibbs sampler described by Garcia-Cortes and Sorensen [32,33]. This requires solving equation (31), which is very sparse, and can be done iteratively. Alternatively, a single-site Gibbs sampler can be used [33]. It can be shown that the full-conditional posterior for σ_g^2 is a scaled inverse chi-square distribution with scale parameter:

$$\hat{S}_g^2 = \frac{\boldsymbol{\epsilon}'\mathbf{A}^{11}\boldsymbol{\epsilon} + S_g^2\nu_g}{\hat{S}_g^2},$$

and degrees of freedom $\hat{S}_g^2 = S_g^2 + n_\epsilon$, where n_ϵ is the number of elements in ϵ [33]. The matrix $\mathbf{A}^{11}$ is very sparse, and thus computing $\boldsymbol{\epsilon}'\mathbf{A}^{11}\boldsymbol{\epsilon}$ is fast.

An alternative sampler

Here, we consider the situation where $\boldsymbol{\epsilon}$ is a "nuisance parameter" and interest is only on inference about marker effects: $\boldsymbol{\alpha}$. The starting point for this sampler is the MME given by equation (23). When $\pi = 0$ in the prior of marker effects of equation (2), conditional on the variance components in the model, the posterior for the location parameters is a normal distribution with mean given by the solution to these MME and covariance matrix given by the the inverse of the left-hand-side of the MME times σ_e^2 [33]. Eliminating $\boldsymbol{\epsilon}$ from (23) results in the following MME:

$$\begin{bmatrix} \mathbf{X}_1^{*'}\mathbf{P}\mathbf{X}_1^* + \mathbf{X}_2^{*'}\mathbf{X}_2^* & \mathbf{X}_1^{*'}\mathbf{P}\mathbf{W}_1 + \mathbf{X}_2^{*'}\mathbf{W}_2 \\ \mathbf{W}_1'\mathbf{P}\mathbf{X}_1^* + \mathbf{W}_2'\mathbf{X}_2^* & \mathbf{W}_1'\mathbf{P}\mathbf{W}_1 + \mathbf{W}_2'\mathbf{W}_2 + \mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}^* \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^{*'}\mathbf{P}\mathbf{y}_1 + \mathbf{X}_2^{*'}\mathbf{y}_2 \\ \mathbf{W}_1'\mathbf{P}\mathbf{y}_1 + \mathbf{W}_2'\mathbf{y}_2 \end{bmatrix}, \quad (32)$$

where $\mathbf{P} = \mathbf{I} - \mathbf{Z}_1\left(\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2}\right)^{-1}\mathbf{Z}_1'$. The solution to these MME for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ are identical to those from equation (23). Furthermore, the inverse elements of the left-hand-side of these MME are identical to the corresponding inverse elements of equation (23). Thus, equation (32) can be used to draw samples from the posterior of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

In BayesA and BayesC with $\pi = 0$, the blocking-Gibbs sampler of [32] can be used to sample all effects jointly. In BayesB, BayesC and BayesCπ, the single-site sampler is more convenient. When the number of observations is larger than the number of effects in the model, equation (13.12) of [33] can be used to compute rhs_i efficiently, which is required to sample the i^{th} location effect from its full-conditional posterior distribution [5,34] as:

$$rhs_i = r_i - \mathbf{C}_i'\boldsymbol{\theta} + \mathbf{C}_{ii}\theta_i, \quad (33)$$

where r_i is the i^{th} element from the right-hand side of equation (32), $\mathbf{C}_i'$ is the i^{th} row from the left-hand side of equation (32), and $\boldsymbol{\theta}$ is the vector of sampled values of the fixed effects and marker effects. Once equation (32) is set up, the time for computing rhs_i as equation (33) does not depend on the number of observations. Thus, computing time for sampling $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by either the blocking-Gibbs sampler or by the single-site sampler, using equation (33) to compute rhs_i, does not depend the number on observations.

Furthermore, when π is close to 1, the sampled value for most marker effects is null. Then, dramatic reductions in

computing time can be achieved as described in the following. Sampling is started with $\boldsymbol{\theta} = \mathbf{0}$, and so initially, the vector $\mathbf{rhs} = \mathbf{r}$. As sampling proceeds and any non-null effect θ_i is sampled, **rhs** is updated as:

$$\mathbf{rhs} = \mathbf{rhs} - \mathbf{C}_i\theta_i, \tag{34}$$

and

$$rhs_i = r_i,$$

where $\mathbf{C}_i$ is the i^{th} column of $\mathbf{C}$. Then, before sampling θ_i, rhs_i would be equal to equation (33). In sampling marker effects, updating **rhs** using equation (34) is the only non-scalar computation, and when π is close to 1, the number of such updates can be very small. In such situations, 40k samples of BayesB can be obtained on a single core in about half an hour, regardless of the number of observations.

The most intensive computation in setting up equation (32) is that of $\mathbf{W}_1'\mathbf{P}\mathbf{W}_1$ and $\mathbf{W}_2'\mathbf{W}_2$. First, we consider computing the matrix of crossproducts: $\mathbf{W}_2'\mathbf{W}_2$. For an arbitrary matrix $\mathbf{S}$ of n rows and k columns, the crossproduct $\mathbf{S}'\mathbf{S}$ can be written as:

$$\mathbf{S}'\mathbf{S} = \sum_i^c \mathbf{S}_i'\mathbf{S}_i, \tag{35}$$

where $\mathbf{S}' = \left[\mathbf{S}_1', \mathbf{S}_2', \ldots, \mathbf{S}_c'\right]$. In equation (35), the crossproducts are independent and can be done in parallel.

Next, we consider computing $\mathbf{W}_1'\mathbf{P}\mathbf{W}_1$. This can be undertaken in two steps. In step 1, the columns of $\mathbf{B} = \mathbf{P}\mathbf{W}_1$ are computed in parallel. Column i of this product can be written as:

$$\mathbf{b}_i = \mathbf{W}_{1i} - \mathbf{Z}_1\left(\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2}\right)^{-1}\mathbf{Z}_1'\mathbf{W}_{1i},$$

where $\mathbf{W}_{1i}$ is the i^{th} column of $\mathbf{W}_1$. The second term in $\mathbf{b}_i$ can be computed efficiently as $\mathbf{Z}_1\mathbf{q}_i$, where $\mathbf{q}_i$ is the solution to the sparse system:

$$\left(\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2}\right)\mathbf{q}_i = \mathbf{Z}_1'\mathbf{W}_{1i}.$$

The Cholesky decomposition of this system is also sufficiently sparse for exact computation and has to be done only once. Once $\mathbf{B}$ is computed, in the second step, the product $\mathbf{W}_1'\mathbf{P}\mathbf{W}_1 = \mathbf{W}_1'\mathbf{B}$ can be computed in parallel, similar to equation (35), as the sum of independent matrix products.

Parallel computations

The Lonestar Linux cluster was used to examine the possible speedup that could be achieved by parallel computing of $\mathbf{S}'\mathbf{S}$. Initially, a single core on this cluster was used to compute $\mathbf{S}'\mathbf{S}$ with the number of rows in $\mathbf{S}$, n, equal to 1 million and the number of columns, k equal to 5000. This computation took 11 669 seconds. If memory was not limiting, computation with $k = 50\,000$ would take 100 times longer because now $\mathbf{S}'\mathbf{S}$ would be 100 times larger than with $k = 5000$. Actual calculations with $k = 50\,000$ were undertaken with 200 nodes. In each node, $\mathbf{S}_i'\mathbf{S}_i$ was computed with $\mathbf{S}_i$ being a slice containing a subset of the 50 000 rows of the $\mathbf{S}$ matrix. MPI [36] was used to compute $\mathbf{S}'\mathbf{S}$ as the sum of these matrices, as in equation (35). The Eigen C++ template library [37], which can exploit the multiple cores within a node, was used to compute $\mathbf{S}_i'\mathbf{S}_i$ within each node. Although each node of the Lonestar cluster has 12 cores, using 8 cores within each node gave the best result: 912 seconds to compute $\mathbf{S}'\mathbf{S}$. This was a speedup of about 1,279 times.

Discussion

Genomic prediction is based either on marker effects models (MEM), where the effects of marker covariates are explicitly included in the model as random effects, or on breeding value models (BVM), where the markers are used to compute the covariance matrix of the breeding values. Although BLUP using these two types of models can be identical [9], computing using the BVM is more efficient when the number of marker covariates is much larger than the number of individuals. Furthermore, the BVM has also been used to combine information from genotyped and non-genotyped individuals to obtain BLUP in a single step (SSBV-BLUP) [11,13,14].

SSBV-BLUP is based on the conditional covariance matrix, $\mathbf{H}$, of $\mathbf{g}$ given both pedigree and observed marker genotypes. When this covariance matrix is written in terms of a single variance component, as in equation (11), a "base correction" is needed to ensure that relationships in $\mathbf{G}$ and $\mathbf{A}$ are expressed relative to the same base or founder population, as explained in detail by Meuwissen et al. [38]. In the MEM (20) presented here, the variance component σ_α^2 for $\boldsymbol{\alpha}$ and σ_g^2 for $\boldsymbol{\epsilon}$ are kept separate. This strategy of keeping σ_α^2 and σ_g^2 separate can also be used in computing $\mathbf{H}$ by defining $\mathbf{G}$ as $\mathbf{M}_2\mathbf{M}_2'\sigma_\alpha^2/\sigma_g^2$ instead of as $\mathbf{M}_2\mathbf{M}_2'/\left[\sum 2p_i(1-p_i)\right]$.

Legarra et al. [11] and Christensen and Lund [14] gave alternative derivations of the matrix $\mathbf{H}$. The derivation by Legarra et al. [11] uses the identity given by equation (7), which was also used here to develop the MEM for single-step Bayesian regression. Here, we reasoned that if $\mathbf{g}$ has a multivariate normal distribution, $\mathbf{g}_2$ and $\boldsymbol{\epsilon}$ would be independent because they are also uncorrelated. Furthermore, we assumed that when conditioning on the observed marker information, the change in the distribution of $\mathbf{g}_1$ results directly from the change in the distribution in $\mathbf{g}_2$. However, one might argue that this assumption is not reasonable because when you condition on the observed

value of $\mathbf{M}_2$, the change in the distribution of $\mathbf{g}_1$ results from the correlated change in the distribution of $\mathbf{M}_1$ and not from the change in the distribution of $\mathbf{g}_2$.

Christensen and Lund [14] did not rely on equation (7) to derive $\mathbf{H}$, but they computed the mean of the missing genotypes conditional on the observed genotypes as:

$$E(\mathbf{M}_1|\mathbf{M}_2) = E(\mathbf{M}_1) + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}[\mathbf{M}_2 - E(\mathbf{M}_2)].$$

However, this is valid only for multivariate normal variables, and multivariate normality is not a good approximation for the distribution of marker covariates. They also used an expression for the conditional variance of the missing genotypes that is valid only for multivariate normal variables. They did not explicitly recognize these approximations in their derivation, but pointed out that the conditional distribution of breeding values of the non-genotyped animals will only be approximately normal. Similar reasoning was also used by de los Campos et al. [39] who clearly showed that conditional on $\mathbf{M}_2$, $\mathbf{g}_1$ has a mixture of scaled-multivariate normal densities.

While we agree that conditional on $\mathbf{M}_2$, $\mathbf{g}_1$ has a mixture of scaled-multivariate normal densities, it must be noted that even the unconditional distribution of $\mathbf{g}_1$ is only approximately normal. Furthermore, in G-BLUP, what is being conditioned on is not the observed value of $\mathbf{M}_2$ but the observed value of $\mathbf{M}_2\mathbf{M}_2'$. To see this, note that there are many different $\mathbf{M}_2$ matrices that result in the same matrix for $\mathbf{G}$, i.e. $\mathbf{M}_2\mathbf{M}_2'$, and thus the same G-BLUP breeding values. In other words, G-BLUP depends on $\mathbf{M}_2$ only through $\mathbf{M}_2\mathbf{M}_2'$, and therefore, all $\mathbf{M}_2$ that result in the same $\mathbf{M}_2\mathbf{M}_2'$ will also result in the same G-BLUP. However, $\mathbf{M}_2\mathbf{M}_2'$ is proportional to the covariance matrix of $\mathbf{g}_2$. Thus, it would be correct to say that in G-BLUP, what is being conditioned on is the covariance matrix of $\mathbf{g}_2$ changing from being proportional to $\mathbf{A}_{22}$ to being proportional to $\mathbf{M}_2\mathbf{M}_2'$. This conditioning may be thought of as a selection process, where the unselected samples of $\mathbf{g}_1$ and $\mathbf{g}_2$ have a multivariate normal distribution with null mean and covariance matrix $\mathbf{A}\sigma_g^2$, while in the selected samples, selection based on $\mathbf{g}_2$ results in its mean still being null but its covariance matrix changing to $\mathbf{M}_2\mathbf{M}_2'\sigma_\alpha^2$. Then, the correlated change in $\mathbf{g}_1$ is given by equation (10), which is a result from Pearson [40] that was also used by Henderson [16] to develop theory for BLUP under a selection model. However, given the selection process that we have described, the distribution of $\mathbf{g}$ may not be multivariate normal.

Hickey et al. [41] imputed genotypes for non-genotyped individuals so that all individuals have genotypes for fitting a MEM. In their case, imputation was undertaken using linkage and linkage disequilibrium information. However, they did not account for any imputation residual. Liu et al. [42] also have described a single-step MEM for combining information from genotyped and non-genotyped animals, which includes a residual polygenic component. Their analysis requires repeated multiplication of $\mathbf{A}_{22}^{-1}$ by a vector, and an efficient algorithm has been developed for this multiplication. In another approach (Theo Meuwissen, personal communication, October 3, 2013), the LDMIP algorithm [43] was used to impute missing genotypes, and for each SNP where the genotype of an individual i was not known with certainty, a random effect with covariance matrix $\mathbf{G}_{LA}$ was fitted to account for the variability that is incompletely explained by the imputed genotype, where $\mathbf{G}_{LA}$ is the linkage analysis based covariance matrix [44]. One of the advantages of this approach is that imputation of genotypes by LDMIP at a locus j uses information from the genotypes of all linked loci, in addition to the genotypes at locus j. The implied imputation in SSBV-BLUP [11-14] and the method presented here, only uses genotypes at the current locus. Furthermore, best linear prediction is used for imputation, which is optimal only for normally distributed variables. In contrast, in LDMIP, conditional probabilities of the missing genotypes, given all observed genotypes at locus j and linked loci, that are computed approximately by iterative peeling and combine both linkage and LD information, are used to impute genotypes. Although the covariance matrix $\mathbf{G}_{LA}$ is easier to justify than the covariance of $\boldsymbol{\epsilon}$ used here and in SSBV-BLUP, normality of the residuals of a single SNP covariate is more difficult to justify than normality of $\boldsymbol{\epsilon}$. Use of mixture genetic models [45] addresses this weakness, but more work is needed to make these analyses efficient for routine use.

The single-step Bayesian regression approach presented here and SSBV-BLUP have the same appealing property that phenotype, genotype and pedigree information are combined in a single-step. Unlike SSBV-BLUP, SSBR is not limited to normally distributed marker effects; SSBR can be used with t-distributed marker effects, as in BayesA, and with mixture models, as in BayesB and BayesCπ. Furthermore, it has the advantage that matrix inversion is not required. However, this comes at the expense of using MCMC methods that are computationally intensive, but these methods have the advantage that computing time and memory requirements increase linearly with the number of observations and number of markers. Thus, as demonstrated here, computing clusters can be used to parallelize and speedup MCMC analyses for routine applications.

Additional file

Additional file 1: Efficient computation of $\hat{\mathbf{M}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2$. It is shown here how the matrix $\hat{\mathbf{M}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2$ of imputed marker covariates can be obtained efficiently by solving an easily formed sparse system of equations.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
RLF conceived the idea for SSBR and developed it in collaboration with DJG and JCMD. DJG proposed the alternative sampler where ϵ is eliminated from the mixed model equations. The manuscript was prepared by RLF with input and suggestions from DJG and JCMD. All authors read and approved the final manuscript.

Acknowledgements
The authors are grateful to Hao Cheng and Claas Heuer for assistance in the investigation of parallel computing to speed up SSBR calculations. This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive grant no. 2012-67015-19420 and by National Institutes of Health grant R01GM099992.

Author details
[1]Department of Animal Science, Iowa State University, 50011 Ames, Iowa, USA. [2]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand.

Received: 5 November 2013 Accepted: 24 June 2014
Published: 22 September 2014

References

1. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: genomic selection in dairy cattle: progress and challenges.** *J Dairy Sci* 2009, **92:**433–443.
2. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: **Invited review: reliability of genomic predictions for north american holstein bulls.** *J Dairy Sci* 2009, **92:**16–24.
3. Wolc A, Stricker C, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, Habier D, Fernando R, Garrick DJ, Lamont SJ, Dekkers JCM: **Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model.** *Genet Sel Evol* 2011, **43:**5.
4. Garrick DJ: **The nature, scope and impact of genomic prediction in beef cattle in the United States.** *Genet Sel Evol* 2011, **43:**17.
5. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819–1829.
6. Nejati-Javaremi A, Smith C, Gibson JP: **Effect of total allelic relationship on accuracy of evaluation and response to selection.** *J Anim Sci* 1997, **75:**1738–1745.
7. Fernando RL: **Genetic evaluation and selection using genotypic, phenotypic and pedigree information.** In *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production: 11-16 June 1998*. Armidale; 1998:329–336.
8. VanRaden PM: **Efficient methods to compute genomic predictions.** *J Dairy Sci* 2008, **91:**4414–4423.
9. Strandén I, Garrick DJ: **Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit.** *J Dairy Sci* 2009, **92:**2971–2975.
10. Saatchi M, McClure MC, McKay SD, Rolf MM, Kim J, Decker JE, Taxis TM, Chapple RH, Ramey HR, Northcutt SL, Bauck S, Woodward B, Dekkers JCM, Fernando RL, Schnabel RD, Garrick DJ, Taylor JF: **Accuracies of genomic breeding values in american angus beef cattle using k-means clustering for cross-validation.** *Genet Sel Evol* 2011, **43:**40.
11. Legarra A, Aguilar I, Misztal I: **A relationship matrix including full pedigree and genomic information.** *J Dairy Sci* 2009, **92:**4656–4663.
12. Misztal I, Legarra A, Aguilar I: **Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information.** *J Dairy Sci* 2009, **92:**4648–4655.
13. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ: **Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of holstein final score.** *J Dairy Sci* 2010, **93:**743–752.
14. Christensen OF, Lund MS: **Genomic prediction when some animals are not genotyped.** *Genet Sel Evol* 2010, **42:**2.
15. Henderson CR, Kempthorne O, Searle SR, Von Krosigk CM: **The estimation of genetic and environmental trends from records subject to culling.** *Biometrics* 1959, **15:**192–218.
16. Henderson CR: **Best linear unbiased estimation and prediction under a selection model.** *Biometrics* 1975, **31:**423–447.
17. Goffinet B: **Selection on selected records.** *Genet Sel Evol* 1983, **15:**91–98.
18. Gianola D, Fernando RL: **Bayesian methods in animal breeding.** *J Anim Sci* 1986, **63:**217–244.
19. Fernando RL, Gianola D: **Statistical inferences in populations undergoing selection or non-random mating.** In *Advances in Statistical Methods for Genetic Improvement of Livestock*. Edited by Gianola D, Hammond K. New York: Springer; 1990:437–457.
20. Im S, Fernando RL, Gianola D: **Likelihood inferences in animal breeding under selection: a missing-data theory view point.** *Genet Sel Evol* 1989, **21:**399–414.
21. Sorensen D, Fernando RL, Gianola D: **Inferring the trajectory of genetic variance in the course of artificial selection.** *Genet Res* 2001, **77:**83–94.
22. Habier D, Fernando RL, Dekkers JCM: **The impact of genetic relationship information on genome-assisted breeding values.** *Genetics* 2007, **177:**2389–2397.
23. Legarra A, Ducrocq V: **Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction.** *J Dairy Sci* 2012, **95:**4629–4645.
24. Faux P, Gengler N, Misztal I: **A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix.** *J Dairy Sci* 2012, **95:**6093–6102.
25. Strandén I, Christensen OF: **Allele coding in genomic evaluation.** *Genet Sel Evol* 2010, **43:**25.
26. Kizilkaya K, Fernando RL, Garrick DJ: **Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes.** *J Anim Sci* 2010, **88:**544–551.
27. Habier D, Fernando RL, Kizilkaya K, Garrick DJ: **Extension of the Bayesian alphabet for genomic selection.** In *Proceedings of the 9th World Congress on Genetics applied to Livestock Production: 1-6 August 2010*. Leipzig; 2010:468. [http://www.kongressband.de/wcgalp2010/assets/pdf/0468.pdf]
28. de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM: **Predicting quantitative traits with regression models for dense molecular markers and pedigree.** *Genetics* 2009, **182:**375–385.
29. Henderson CR: *Applications of Linear Models in Animal Breeding*. Edited by Shaeffer LR. Guelph: University of Guelph; 1984.
30. Fernando RL, Habier D, Stricker C, Dekkers JCM, Totir LR: **Genomic selection.** *Acta Agric Scand* 2007, **57:**192–195.
31. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R: **Additive genetic variability and the bayesian alphabet.** *Genetics* 2009, **183:**347–363.
32. Garcia-Cortes LA, Sorensen D: **On a multivariate implementation of the Gibbs sampler.** *Genet Sel Evol* 1996, **28:**121–126.
33. Sorensen DA, Gianola D: *Likelihood, Bayesian,and MCMC Methods in Quantitative Genetics*. New York: Springer; 2002.
34. Fernando R, Garrick D: **Bayesian methods applied to GWAS.** In *Genome-Wide Association Studies and Genomic Prediction*. Edited by Gondro C, van der Werf J, Hayes B. New York: Humana Press; 2013.
35. Legarra A, Misztal I: **Technical note: computing strategies in genome-wide selection.** *J Dairy Sci* 2007, **91:**360–366.
36. Message Passing Interface Forum: *MPI: A Message-Passing Interface Standard, Version 3.0*. Tennessee: University of Tennessee; 2012.
37. Guennebaud G, Jacob B: **Eigen v3.** 2010. [http://eigen.tuxfamily.org]
38. Meuwissen THE, Luan T, Woolliams JA: **The unified approach to the use of genomic and pedigree information in genomic evaluations revisited.** *J Anim Breed Genet* 2011, **128:**429–439.
39. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP: **Whole-genome regression and prediction methods applied to plant and animal breeding.** *Genetics* 2013, **193:**327–345.
40. Pearson K: **Mathematical contributions to the theory of evolution. XI. on the influence of natural selection on the variability and correlation of organs.** *Philo Roy Soc* 1903, **200:**1–66.

41. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA: **A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation.** *Genet Sel Evol* 2012, **44:**9.
42. Liu R, Goddard ME, Reinhardt F, Reents R: **Computing strategies for a single step SNP model with an across country reference population.** In *Proceedings of the 2013 Annual Meeting of the European Federation of Animal Science: 26-30 August 2013*. Nantes; 2013:452.
43. Meuwissen T, Goddard M: **The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data.** *Genetics* 2010, **185:**1441–1449.
44. Fernando RL, Grossman M: **Marker assisted selection using best linear unbiased prediction.** *Genet Sel Evol* 1989, **21:**467–477.
45. Habier D, Totir LR, Fernando RL: **A two-stage approximation for analysis of mixture genetic models in large pedigrees.** *Genetics* 2010, **185:**655–670.

doi:10.1186/1297-9686-46-50
Cite this article as: Fernando *et al.*: A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics Selection Evolution* 2014 **46**:50.