Heredity (2020) 124:274–287
https://doi.org/10.1038/s41437-019-0273-4

the geneticssociety

ARTICLE

Multi-trait single-step genomic prediction accounting for heterogeneous (co)variances over the genome

Emre Karaman[1] · Mogens S. Lund[1] · Guosheng Su[1]

Received: 3 April 2019 / Revised: 5 September 2019 / Accepted: 6 September 2019 / Published online: 22 October 2019

Abstract

Widely used genomic prediction models may not properly account for heterogeneous (co)variance structure across the genome. Models such as BayesA and BayesB assume locus-specific variance, which are highly influenced by the prior for (co)variance of single nucleotide polymorphism (SNP) effect, regardless of the size of data. Models such as BayesC or GBLUP assume a common (co)variance for a proportion (BayesC) or all (GBLUP) of the SNP effects. In this study, we propose a multi-trait Bayesian whole genome regression method (BayesN0), which is based on grouping a number of predefined SNPs to account for heterogeneous (co)variance structure across the genome. This model was also implemented in single-step Bayesian regression (ssBayesN0). For practical implementation, we considered multi-trait single-step SNPBLUP models, using (co)variance estimates from BayesN0 or ssBayesN0. Genotype data were simulated using haplotypes on first five chromosomes of 2200 Danish Holstein cattle, and phenotypes were simulated for two traits with heritabilities 0.1 or 0.4, assuming 200 quantitative trait loci (QTL). We compared prediction accuracy from different prediction models and different region sizes (one SNP, 100 SNPs, one chromosome or whole genome). In general, highest accuracies were obtained when 100 adjacent SNPs were grouped together. The ssBayesN0 improved accuracies over BayesN0, and using (co)variance estimates from ssBayesN0 generally yielded higher accuracies than using (co)variance estimates from BayesN0, for the 100 SNPs region size. Our results suggest that it could be a good strategy to estimate (co)variance components from ssBayesN0, and then to use those estimates in genomic prediction using multi-trait single-step SNPBLUP, in routine genomic evaluations.

Background

Genomic selection was pioneered by the study of Meuwissen et al. (2001), and is rapidly becoming the state-of-the-art genetic selection methodology in many breeding programs around the world. The models proposed by Meuwissen et al. (2001) include a BLUP model, where the variances of single nucleotide polymorphism (SNP) effects are assumed to be the same for all SNPs (SNPBLUP), or specific to each SNP (BayesA and BayesB). Under a series of assumptions, the SNPBLUP model is equivalent to a mixed linear model, GBLUP (Habier et al. 2007), which uses a relationship matrix (**G**) computed from genetic markers (Nejati-Javaremi et al. 1997) to model covariances between individuals' genetic effects (Stranden and Garrick 2009). This equivalency resulted in a widespread adoption of genomic prediction in genetic evaluations, because only an extra step of computation of **G** and its inverse is required for the traditional mixed model equations (Henderson 1984) used in animal breeding (Karaman et al. 2016). Moreover, it also allows all extensions of BLUP methodology, such as multiple-trait, random regression, or repeated measures to be easily implemented in genomic evaluations (Tiezzi and Maltecca 2015). The GBLUP model has been widely used to predict breeding values in animal species, such as cattle (Luan et al. 2009; Su et al. 2012b), pig (Lukić et al. 2015), sheep (Daetwyler et al. 2010a) and fish (Ødegård et al. 2014; Tsai et al. 2016), and accuracies from GBLUP were reported to be higher than those from traditional pedigree-based BLUP.

Supplementary information The online version of this article (https://doi.org/10.1038/s41437-019-0273-4) contains supplementary material, which is available to authorized users.

✉ Emre Karaman
emre@mbg.au.dk

[1] Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark

Although widely used in genomic evaluations, these BLUP-based genomic prediction models have some drawbacks. First, they ignore the fact that a large proportion of the SNPs may not have any influence on the trait of interest. Second, different loci or genomic regions may have rather different variances. The two models of Meuwissen et al. (2001), BayesA and BayesB, were proposed to overcome such drawbacks. Assuming SNP-specific variances, BayesA fits each of SNPs, while BayesB fits approximately 1-π of the SNPs, where π is the percentage of SNPs which have no influence on the trait of interest. When $\pi = 0$, BayesB is equivalent to BayesA. As pointed out by Gianola et al. (2009), both models are problematic as full conditional posteriors of the SNP-specific variances have only one additional degree of freedom compared to their priors regardless of the amount of data available. A simpler model that similarly fits approximately 1-π of the SNPs, but with a common variance, BayesC, was also proposed (Meuwissen 2009; Kizilkaya et al. 2010).

Zeng et al. (2016) introduced a Bayesian partitioned regression model for genomic prediction, which involves the selection of genome regions followed by the selection of SNPs within those selected regions. The model fits approximately $1 - \Pi$ of the regions assuming region-specific variances, and $1 - \pi_s$ of the SNPs within the region s assuming a common variance for the SNPs in the region. Referring to this "nested" variable selection structure of the model, it was termed as BayesN. The special case of the partitioned regression model of Zeng et al. (2016), i.e., BayesN with $\Pi = \pi_s = 0$ (hereafter, BayesN0), is equivalent to BayesA or GBLUP when a fixed region size is set at one SNP or the whole genome, respectively. We hypothesize that, at any other region size, but these two extreme sizes of genome regions, higher prediction accuracies can be obtained using BayesN0. Although it ignores the fact that a proportion of the genome regions, and therefore a proportion of the SNPs, may not have any influence on the trait of interest, prediction accuracy may increase compared to BayesA by benefiting from the increase in the accuracy in estimation of SNP variances, and compared to BLUP-based models by allowing SNPs in different regions to have different variances. Partitioning of the covariate matrix of marker genotypes, **M**, or in other words, assigning priors to genome regions rather than individual SNPs, was shown to influence the accuracy of genomic predictions (Brøndum et al. 2012; Gebreyesus et al. 2017; Karaman et al. 2018).

Many important traits in animal breeding have genetic correlations in varying sizes with one or more traits, and therefore, measurements of such correlated traits carry information for the genetic values of others. Several multi-trait models have been proposed for genomic prediction (Calus and Veerkamp 2011; Jia and Jannink 2012; Hayashi and Iwata 2013; Gebreyesus et al. 2017; Cheng et al. 2018b), and simulations have shown that genomic prediction accuracies from multi-trait models are superior to those from single-trait models (Calus and Veerkamp 2011; Jia and Jannink 2012; Guo et al. 2014; Karaman et al. 2018). Multi-trait genetic evaluation rely on the genetic association between the traits through the genetic variance and covariance structure. Models used for genomic prediction, therefore, should properly account for the makeup of these genetic (co)variance components to obtain the highest accuracy of prediction. When only a few genome regions explain a considerable amount of the variances and/or covariance in a two-trait analysis, models that account for the heterogeneous correlation structure over the genome may have advantages over the methods that assumes a constant correlation over the genome (Gebreyesus et al. 2017; Karaman et al. 2018).

The GBLUP model was extended to utilize all phenotypic, pedigree and genotypic information simultaneously, including phenotypic information on non-genotyped individuals, and termed as single-step GBLUP (ssGBLUP) (Christensen and Lund 2010; Aguilar et al. 2010). In ssGBLUP, the pedigree-based relationship matrix **A** and the genomic relationship matrix **G** are combined into a single matrix **H**. As for GBLUP, only an extra step for computation of **H** and its inverse is required for the traditional mixed model equations used in animal breeding (Misztal and Legarra 2017). However, ssGBLUP also suffers from the same drawbacks of GBLUP.

Fernando et al. (2014) proposed a class of single-step models, which not only unifies all available information as ssGBLUP does, but also accommodates any Bayesian whole genome regression model. This yields models of, for instance, ssBayesA or ssBayesN0, referring to the Bayesian whole genome regression model used in the single-step analysis. However, such an approach requires that all unknowns of the model to be estimated using Markov-chain Monte Carlo techniques which may be computationally infeasible especially in routine genomic evaluations. In genomic predictions using weighted GBLUP, it was shown that the use of the same SNP variances over a few years does not reduce prediction accuracy (Su et al. 2014). Indeed, in routine evaluations, variance components are not updated for each round of evaluation, because they are expected to be relatively consistent over time (Calus et al. 2014). An alternative to the fully Bayesian approach in Fernando et al. (2014) could be a strategy, where all necessary parameters are estimated using a Bayesian whole genome regression model first, and mixed model equations are then solved given the "known" values of the variance components, leading to a single-step SNPBLUP (ssSNPBLUP) model.

The aim of this study was three-fold: (i) to introduce a multi-trait whole genome regression model that allows heterogeneous (co)variances, (ii) to compare accuracies from single- and multi-trait genomic prediction, and (iii) to investigate the use of region-specific estimates of (co)variances in genomic predictions using ssSNPBLUP.

Material and methods

Data sets and simulations

The genotype data were simulated for five generations (Gen1−Gen5) based on real haplotypes of 2200 Holsteins (Gen0), as described in Karaman et al. (2018). At each generation, the number of males and females were kept constant at 200 and 2000, respectively, and the mating ratio was 1:10. Mating was completely at random, and selection was not considered. Each sire was mated twice with one of the ten dams to keep the population size at 2200 at each generation. Only the single nucleotide polymorphisms (SNPs) (11,154) located on first five chromosomes were considered.

Phenotypic values of the two traits were simulated to have heritabilities of 0.1 and 0.4, which represents low (L) and high (H) heritability traits, respectively. Total number of quantitative trait loci (QTL) was set at 200, which were randomly selected from the SNP set, ensuring that the average minor allele frequency (MAF) of QTL is 0.15 (Karaman et al. 2018). The criterion for the MAF of the QTL was based on the assumption that they in general have relatively low MAF (Goddard and Hayes 2009; Kemper and Goddard 2012). The QTL were randomly assigned into three groups according to their causal relationships with the traits. This was done by assuming a percentage of the total QTL (82%) had pleiotropic effects on two traits, while one half of the remaining QTL had effect on one trait, and one half on the other trait.

Two scenarios, G9 and N5, were considered in terms of the distribution of QTL effects and correlations for the effect of pleiotropic QTL. In the scenario G9, the effects of the pleiotropic QTL were achieved by simulating two correlated gamma variables (Dvorkin 2012) with marginal distributions of $G(0.4, 1.66)$, and a correlation of 0.9. The 78% of those QTL were assigned to a correlation between effects on two traits of 0.9, and 22% of −0.9 randomly. The correlation group of −0.9 was achieved by switching the sign of QTL effect for one of the traits at random. The QTL effects, which were assumed to have a correlation of 0.9, were assigned a negative or positive sign at random for both traits. In the second scenario, scenario N5, effects of all pleiotropic QTL were simulated from a bivariate normal distribution with a correlation of 0.5. Although fluctuated across the replicates, all scenarios lead to genetic correlations of about 0.45 at Gen0. The QTL SNPs were excluded from the final data set of SNP for the analysis. Random residual effects were sampled from $N\left(\mathbf{0}, \begin{bmatrix} \mathbf{I}\sigma^2_{e_\mathrm{L}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma^2_{e_\mathrm{H}} \end{bmatrix}\right)$, where the sizes of $\sigma^2_{e_\mathrm{L}}$ and $\sigma^2_{e_\mathrm{H}}$ were determined according to heritabilities of 0.1 and 0.4, respectively.

Table 1 Number of animals with genotype and phenotype in each generation

Generation	G	G&P	P	Total
Gen0	–	–	–	2200
Gen1	–	–	–	2200
Gen2	–	–	–	2200
Gen3	200 (M)	500 (F)	1500 (F)	2200
Gen4	200 (M)	500 (F)	1500 (F)	2200
Gen5	500	–	–	2200

G genotype, *P* phenotype, *M* male, *F* female

Final data (see Table 1) were created by masking genotypes and/or phenotypes of the animals as follows. For generations 3 and 4, it was assumed that males had no phenotypes, but genotypes, while all females had phenotypes, and some fraction of them had also genotypes. Those genotyped females were selected completely at random. Generation 5 was used as validation population, where 500 randomly selected animals were assumed to be genotyped. Pedigree was traced back to Gen0. Animals had phenotypes on both traits, or none of them. In total, 20 replicates were generated.

Models and methods

A novel multi-trait Bayesian whole genome regression model (BayesN0), single-step SNPBLUP and single-step Bayesian regression models introduced by Fernando et al. (2014) were compared for multi-trait genomic prediction. Single-trait analysis were also performed, but neither the models nor their theory were given in this paper, as the models are special cases of their multi-trait counterparts. In this section, we followed the notation in Fernando et al. (2014) as closely as possible.

Basic multi-trait model

A multi-trait mixed model including only general means as fixed effects and marker effects as random effects can be written as

$$\begin{bmatrix} \mathbf{y}_\mathrm{L} \\ \mathbf{y}_\mathrm{H} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_\mathrm{L} & 0 \\ 0 & \mathbf{1}_\mathrm{H} \end{bmatrix} \begin{bmatrix} \mu_\mathrm{L} \\ \mu_\mathrm{H} \end{bmatrix} + \begin{bmatrix} \mathbf{M}_\mathrm{L} & 0 \\ 0 & \mathbf{M}_\mathrm{H} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_\mathrm{L} \\ \boldsymbol{\alpha}_\mathrm{H} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_\mathrm{L} \\ \mathbf{e}_\mathrm{H} \end{bmatrix}, \tag{1}$$

where $\mathbf{y}_\mathrm{L}$ and $\mathbf{y}_\mathrm{H}$ are the vectors of phenotypes, $\mathbf{1}$ are vectors of ones, μ_L and μ_H are general means, $\mathbf{M}_\mathrm{L}$ and $\mathbf{M}_\mathrm{H}$ are the matrices of genotypes for k markers, $\boldsymbol{\alpha}_\mathrm{L}$ and $\boldsymbol{\alpha}_\mathrm{H}$ are the vectors of marker effects, and $\mathbf{e}_{\mathrm{L}}$ and $\mathbf{e}_{\mathrm{H}}$ are the vectors of random residual effects, for traits "L" and "H", respectively. In our simulations, animals had records for both traits or none of them. Therefore, $\mathbf{M}_{\mathrm{L}} = \mathbf{M}_{\mathrm{H}}$, and these matrices will be denoted as $\mathbf{M}$ hereinafter, to simplify the demonstration. Residuals, $\mathbf{e}' = \left[\mathbf{e}_{\mathrm{L}}^{'}, \mathbf{e}_{\mathrm{H}}^{'}\right]$, are typically assumed to follow a normal distribution, $\mathbf{e} \mid \mathbf{R}_0 \sim N(\mathbf{0}, \mathbf{R}_0 \otimes \mathbf{I})$, where $\mathbf{R}_0 = \begin{bmatrix} \sigma^2_{e_{\mathrm{L}}} & \sigma_{e_{\mathrm{LH}}} \\ \sigma_{e_{\mathrm{HL}}} & \sigma^2_{e_{\mathrm{H}}} \end{bmatrix}$, and $\mathbf{I}$ is an identity matrix.

Multi-trait Bayesian partitioned regression (BayesN0)

The columns of $\mathbf{M}$ and vector $\boldsymbol{\alpha}$ given in Eq. (1) can be divided into S subsets in a conformable manner:

$$\begin{bmatrix} \mathbf{y}_{\mathrm{L}} \\ \mathbf{y}_{\mathrm{H}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{\mathrm{L}} & 0 \\ 0 & \mathbf{1}_{\mathrm{H}} \end{bmatrix} \begin{bmatrix} \mu_{\mathrm{L}} \\ \mu_{\mathrm{H}} \end{bmatrix} + \begin{bmatrix} \mathbf{M}_1 \ldots \mathbf{M}_S & 0 \\ 0 & \mathbf{M}_1 \ldots \mathbf{M}_S \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_{\mathrm{L},1} \\ \vdots \\ \boldsymbol{\alpha}_{\mathrm{L},S} \\ \boldsymbol{\alpha}_{\mathrm{H},1} \\ \vdots \\ \boldsymbol{\alpha}_{\mathrm{H},S} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_{\mathrm{L}} \\ \mathbf{e}_{\mathrm{H}} \end{bmatrix},$$

where $\mathbf{y} = \begin{bmatrix} \mathbf{y}_{\mathrm{L}} \\ \mathbf{y}_{\mathrm{H}} \end{bmatrix}$ involves the phenotypes of genotyped individuals only, $\mathbf{M}_1, \ldots, \mathbf{M}_S$ are genotype matrices regarding genomic regions, and $\boldsymbol{\alpha}_{t,1}, \ldots, \boldsymbol{\alpha}_{t,S}$ ($t = $ L, H for low and high heritability traits, respectively) are vectors of SNP effects for corresponding genomic regions. We assume that all SNPs j ($j = 1, \ldots, k_s$) in the same genomic region s ($s = 1, \ldots, S$) have the same (co)variance for the two traits:

$$\mathrm{var}\left(\boldsymbol{\alpha}_{sj}\right) = \mathrm{var}\begin{bmatrix} \alpha_{\mathrm{L},sj} \\ \alpha_{\mathrm{H},sj} \end{bmatrix} = \mathbf{B}_s = \begin{bmatrix} \sigma^2_{\alpha_{\mathrm{L},s}} & \sigma_{\alpha_{\mathrm{LH},s}} \\ \sigma_{\alpha_{\mathrm{HL},s}} & \sigma^2_{\alpha_{\mathrm{H},s}} \end{bmatrix}.$$

Likelihood of the model is given as:

$$p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{B}, \mathbf{R}) \propto |\mathbf{R}|^{-\frac{1}{2}}\exp\Big\{-\tfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu} - \mathbf{M}_1\boldsymbol{\alpha}_1 - \ldots \\ -\mathbf{M}_S\alpha_S)^{'}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu} - \mathbf{M}_1\boldsymbol{\alpha}_1 - \ldots\ - \mathbf{M}_S\alpha_S)\Big\}$$

where $\mathbf{X} = \begin{bmatrix} \mathbf{1}_{\mathrm{L}} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{\mathrm{H}} \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \mu_{\mathrm{L}} \\ \mu_{\mathrm{H}} \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} \mathbf{B}_{\mathrm{L}} & \mathbf{B}_{\mathrm{LH}} \\ \mathbf{B}_{\mathrm{HL}} & \mathbf{B}_{\mathrm{H}} \end{bmatrix}$ with $\mathbf{B}_i$ being diagonal matrices consisting of SNP variances ($\mathbf{B}_{\mathrm{L}}$ and $\mathbf{B}_{\mathrm{H}}$) or covariances ($\mathbf{B}_{\mathrm{LH}} = \mathbf{B}_{\mathrm{HL}}$), and $\mathbf{R} = \mathbf{R}_{\mathbf{0}} \otimes \mathbf{I}$. The vector of fixed effects, $\boldsymbol{\mu}$, were assigned a flat prior, and other parameters of the model were assigned a normal or an inverse Wishart (IW) prior for conjugacy:

$$\boldsymbol{\alpha}_{sj} \mid \mathbf{B}_s \sim N(\mathbf{0}, \mathbf{B}_{\mathbf{s}})$$

$$\mathbf{e} \mid \mathbf{R}_0 \sim N(\mathbf{0}, \mathbf{R}_0 \otimes \mathbf{I})$$

$$\mathbf{B}_s \mid v_B, \mathbf{V}_B \sim IW(v_B, \mathbf{V}_B)$$

$$\mathbf{R}_0 \mid v_R, \mathbf{V}_R \sim IW(v_R, \mathbf{V}_R).$$

Full conditional distributions of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}_{sj}$, $\mathbf{B}_s$, and $\mathbf{R}_0$ can be obtained after some algebra:

$$p(\boldsymbol{\mu} \mid .) \sim N\left[\left(\mathbf{X}'\mathbf{R}^{-\mathbf{1}}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{R}^{-\mathbf{1}}\mathbf{y}^*, \left(\mathbf{X}'\mathbf{R}^{-\mathbf{1}}\mathbf{X}\right)^{-\mathbf{1}}\right]$$

$$p(\boldsymbol{\alpha}_{sj} \mid .) \sim N\left[\left(\mathbf{M}_{\mathbf{j}}^{*'}\mathbf{R}^{-\mathbf{1}}\mathbf{M}_{\mathbf{j}}^* + \mathbf{B}_{\mathbf{s}}^{-\mathbf{1}}\right)^{-1}\mathbf{M}_{\mathbf{j}}^{*'}\mathbf{R}^{-1}\mathbf{y}^*, \left(\mathbf{M}_{\mathbf{j}}^{*'}\mathbf{R}^{-\mathbf{1}}\mathbf{M}_{\mathbf{j}}^* + \mathbf{B}_{\mathbf{s}}^{-\mathbf{1}}\right)^{-\mathbf{1}}\right]$$

$$p(\mathbf{B}_s \mid .) \sim IW[v_B + k_s, (\mathbf{S}_{B_s} + \mathbf{V}_B)]$$

$$p(\mathbf{R}_0 \mid .) \sim IW[v_R + n, (\mathbf{S}_R + \mathbf{V}_R)],$$

where, "." stands for all other parameters and $\mathbf{y}^*$, $\mathbf{y}^*$ is the vector of phenotypes corrected for all other effects, $\mathbf{M}_j^* = \begin{bmatrix} \mathbf{m}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{m}_j \end{bmatrix}$, $\mathbf{S}_{B_s} = \sum_{j=1}^{k_s} \boldsymbol{\alpha}_{sj}\boldsymbol{\alpha}_{sj}^{'}$ and $\mathbf{S}_R = \sum_{i=1}^{n} \mathbf{e}_i\mathbf{e}_i^{'}$. This multi-trait whole genome regression model was referred to as multi-trait BayesN0 throughout this paper, as it is an extension of a particular form of partitioned regression model (BayesN) introduced by Zeng et al. (2016), to multi-trait case. Note that when the size of region is fixed at one SNP or whole genome, model becomes equivalent to multi-trait BayesA or GBLUP, respectively.

Multi-trait single-step SNPBLUP

In the following expressions, n stands for the non-genotyped animals, and g stands for the genotyped animals. Note that in our simulations, animals had records for both traits or none of them. In a multi-trait single-step SNPBLUP (ssSNPBLUP) analysis, the phenotypes are modeled as (Fernando et al. 2014):

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\mu}^* + \mathbf{W}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\epsilon} + \mathbf{e}\,, \qquad (2)$$

where $\mathbf{y} = \begin{bmatrix} \mathbf{y}_{\mathrm{L}} \\ \mathbf{y}_{\mathrm{H}} \end{bmatrix}$ is the vector of phenotypes for genotyped and non-genotyped individuals, $\boldsymbol{\mu}^* = \begin{bmatrix} \mu_{\mathrm{L}} \\ \mu_{\mathrm{g,L}} \\ \mu_{\mathrm{H}} \\ \mu_{\mathrm{g,H}} \end{bmatrix}$, μ_{L} and μ_{H} are the overall means of the two traits, $\mu_{\mathrm{g,L}}$ and $\mu_{\mathrm{g,H}}$ are the differences between breeding values of genotyped and non-genotyped animals for the two traits, $\mathbf{X}^* = \begin{bmatrix} \mathbf{X}_{\mathrm{L}}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{\mathrm{H}}^* \end{bmatrix}$ with $\mathbf{X}_{\mathrm{L}}^* = \mathbf{X}_{\mathrm{H}}^* = \begin{bmatrix} \mathbf{1} & -\mathbf{Z}_{\mathrm{n}}\mathbf{A}_{\mathrm{ng}}\mathbf{A}_{\mathrm{gg}}^{-1}\mathbf{1} \\ \mathbf{1} & -\mathbf{Z}_{\mathrm{g}}\mathbf{1} \end{bmatrix}$, $\mathbf{W} = \begin{bmatrix} \mathbf{Z}_{\mathrm{L}} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{\mathrm{H}} \end{bmatrix}$ $\begin{bmatrix} \mathbf{M}_{\mathrm{L}} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{\mathrm{H}} \end{bmatrix}$ with $\mathbf{Z}_{\mathrm{L}} = \mathbf{Z}_{\mathrm{H}} = \begin{bmatrix} \mathbf{Z}_{\mathrm{n}} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{\mathrm{g}} \end{bmatrix}$ and

$\mathbf{M}_{\mathrm{L}} = \mathbf{M}_{\mathrm{H}} = \begin{bmatrix} \widehat{\mathbf{M}}_{\mathrm{n}} \\ \mathbf{M}_{\mathrm{g}} \end{bmatrix}$, $\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_{\mathrm{L}} \\ \boldsymbol{\alpha}_{\mathrm{H}} \end{bmatrix}$. $\mathbf{Z}_{\mathrm{n}}$ and $\mathbf{Z}_{\mathrm{g}}$ are incidence matrices relating breeding values of non-genotyped and genotyped animals to their phenotypes, $\widehat{\mathbf{M}}_{\mathrm{n}}$ and $\mathbf{M}_{\mathrm{g}}$ are matrices of imputed and observed genotypes for non-genotyped and genotyped animals, respectively, $\boldsymbol{\alpha}_{\mathrm{L}}$ and $\boldsymbol{\alpha}_{\mathrm{H}}$ are the vectors of allele substitution effects. The $\mathbf{U} = \begin{bmatrix} \mathbf{U}_{\mathrm{L}} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_{\mathrm{H}} \end{bmatrix}$ with $\mathbf{U}_{\mathrm{L}} = \mathbf{U}_{\mathrm{H}} = \begin{bmatrix} \mathbf{Z}_{\mathrm{n}} \\ \mathbf{0} \end{bmatrix}$ and $\boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_{\mathrm{L}} \\ \boldsymbol{\epsilon}_{\mathrm{H}} \end{bmatrix}$, where $\boldsymbol{\epsilon}_{\mathrm{L}}$ and $\boldsymbol{\epsilon}_{\mathrm{H}}$ are the vectors of imputation residuals. The $\mathbf{e}$ is a vector of random residual effects assumed to follow $\mathbf{e} \mid \mathbf{R}_0 \sim N(\mathbf{0},\ \mathbf{R}_0 \otimes \mathbf{I})$, where $\mathbf{R}_0 = \begin{bmatrix} \sigma^2_{e_{\mathrm{L}}} & \sigma_{e_{\mathrm{LH}}} \\ \sigma_{e_{\mathrm{HL}}} & \sigma^2_{e_{\mathrm{H}}} \end{bmatrix}$, and $\mathbf{I}$ is an identity matrix. Vector of $\boldsymbol{\alpha}$ is assumed to follow $\boldsymbol{\alpha} \mid \mathbf{B} \sim N(\mathbf{0},\ \mathbf{B})$ with $\mathbf{B} = \begin{bmatrix} \mathbf{B}_{\mathrm{L}} & \mathbf{B}_{\mathrm{LH}} \\ \mathbf{B}_{\mathrm{HL}} & \mathbf{B}_{\mathrm{H}} \end{bmatrix}$, where $\mathbf{B}_i$ are diagonal matrices consisting of SNP variances ($\mathbf{B}_{\mathrm{L}}$ and $\mathbf{B}_{\mathrm{H}}$) or covariances ($\mathbf{B}_{\mathrm{LH}} = \mathbf{B}_{\mathrm{HL}}$). Vector of $\boldsymbol{\epsilon}$ is assumed to follow $\boldsymbol{\epsilon} \mid \mathbf{G}_0, \mathbf{A} \sim N(\mathbf{0}, \mathbf{G}_0 \otimes \mathbf{A})$, where $\mathbf{G}_0$ is the additive genetic (co)variance matrix. The $\mathbf{A}_{\mathrm{ng}}$, $\mathbf{A}_{\mathrm{gg}}$ and $\mathbf{A}_{\mathrm{nn}}$ are submatrices of the pedigree-based relationship matrix, $\mathbf{A}$, corresponding to the relationships between non-genotyped and genotyped individuals, among the genotyped individuals, and among the non-genotyped individuals, respectively. The matrix of imputed genotypes, $\widehat{\mathbf{M}}_{\mathrm{n}}$, is obtained with $\mathbf{A}_{\mathrm{ng}}\mathbf{A}_{\mathrm{gg}}^{-1}\mathbf{M}_{\mathrm{g}}$ (Fernando et al. 2014).

The mixed model equations corresponding to the model in Eq. (2) is as follows.

$$\begin{bmatrix} \mathbf{X}^{*\prime}\mathbf{R}^{-1}\mathbf{X}^{*} & \mathbf{X}^{*\prime}\mathbf{R}^{-1}\mathbf{W} & \mathbf{X}_{\mathrm{n}}^{*\prime}\mathbf{R}^{-1}\mathbf{U}_{\mathrm{n}} \\ \mathbf{W}^{\prime}\mathbf{R}^{-1}\mathbf{X}^{*} & \mathbf{W}^{\prime}\mathbf{R}^{-1}\mathbf{W} + \mathbf{B}^{-1} & \mathbf{W}_{\mathrm{n}}^{\prime}\mathbf{R}^{-1}\mathbf{U}_{\mathrm{n}} \\ \mathbf{U}_{\mathrm{n}}^{\prime}\mathbf{R}^{-1}\mathbf{X}^{*} & \mathbf{U}_{\mathrm{n}}^{\prime}\mathbf{R}^{-1}\mathbf{W} & \mathbf{U}_{\mathrm{n}}^{\prime}\mathbf{R}^{-1}\mathbf{U}_{\mathrm{n}} + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-\mathrm{nn}} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\mu}} \\ \widehat{\boldsymbol{\alpha}} \\ \widehat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{*\prime}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}^{\prime}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{U}_{\mathrm{n}}^{\prime}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{3}$$

The $\mathbf{A}^{-\mathrm{nn}}$ is the part of the inverse of pedigree-based relationship matrix, $\mathbf{A}$, corresponding to the non-genotyped individuals, and $\mathbf{R} = \mathbf{R}_0 \otimes \mathbf{I}$.

Multi-trait single-step BayesN0 (ssBayesN0)

The single-step SNPBLUP requires the estimation of (co) variance components, and then use of these in mixed model equations to estimate breeding values. In contrast, Bayesian approach can be used to obtain the vector of fixed and random effect estimates, $[\widehat{\mu}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\epsilon}}]^{\prime}$, the genetic and residual variance components, and the SNP (co) variances simultaneously, as in the original paper of Fernando et al. (2014). In principle, any Bayesian whole genome regression model can be incorporated in this single-step model, and BayesN0 was used here (ssBayesN0). Likelihood of the ssBayesN0 model is given as:

$$p(\mathbf{y} \mid \boldsymbol{\mu}^{*}, \boldsymbol{\alpha}, \mathbf{B}, \mathbf{R}) \propto |\mathbf{R}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{y} - \mathbf{X}^{*}\boldsymbol{\mu}^{*} - \mathbf{W}_1\boldsymbol{\alpha}_1 - \ldots\ - \mathbf{W}_S\boldsymbol{\alpha}_S - \mathbf{U}\boldsymbol{\epsilon})^{\prime}\mathbf{R}^{-1} \right.$$
$$\left.(\mathbf{y} - \mathbf{X}^{*}\boldsymbol{\mu}^{*} - \mathbf{W}_1\boldsymbol{\alpha}_1 - \cdots - \mathbf{W}_S\boldsymbol{\alpha}_S - \mathbf{U}\boldsymbol{\epsilon})\right\},$$

where matrices and parameters are as specified earlier. A flat prior was assumed for $\boldsymbol{\mu}^*$. Priors for $\boldsymbol{\alpha}$, $\mathbf{e}$, $\mathbf{B}_s$ and $\mathbf{R}_0$ were the same as in BayesN0. A multivariate normal prior, $\boldsymbol{\epsilon} \mid \mathbf{G}_0, \mathbf{A} \sim N(\mathbf{0}, \mathbf{G}_0 \otimes \mathbf{A})$, was assumed for the vector of $\boldsymbol{\epsilon}$, and $\mathbf{G}_0$ was assigned an inverse Wishart prior, $\mathbf{G}_0 \mid \nu_{\mathrm{G}}, \mathbf{V}_{\mathrm{G}} \sim IW(\nu_{\mathrm{G}},\ \mathbf{V}_{\mathrm{G}})$. Full conditional distributions of $\boldsymbol{\mu}^*$, $\boldsymbol{\alpha}_{sj}$, $\mathbf{B}_s$, $\boldsymbol{\epsilon}$, $\mathbf{G}_0$, $\mathbf{R}_0$ can be obtained after some algebra:

$$p(\boldsymbol{\mu}^{*} \mid .) \sim N\left[\left(\mathbf{X}^{\star\prime}\mathbf{R}^{-1}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{\prime}\mathbf{R}^{-1}\mathbf{y}^{*},\ \left(\mathbf{X}^{\star\prime}\mathbf{R}^{-1}\mathbf{X}^{*}\right)^{-1}\right]$$

$$p(\boldsymbol{\alpha}_{sj} \mid .) \sim N\left[\left(\mathbf{W}_j^{*\prime}\mathbf{R}^{-1}\mathbf{W}_j^{*} + \mathbf{B}_s^{-1}\right)^{-1}\mathbf{W}_j^{*\prime}\mathbf{R}^{-1}\mathbf{y}^{*},\ \left(\mathbf{W}_j^{*\prime}\mathbf{R}^{-1}\mathbf{W}_j^{*} + \mathbf{B}_s^{-1}\right)^{-1}\right]$$

$$p(\mathbf{B}_s \mid .) \sim IW[\nu_B + k_s, (\mathbf{S}_{B_s} + \mathbf{V}_B)]$$

$$p(\boldsymbol{\epsilon}|.) \sim N\left[\left(\mathbf{U}_{\mathrm{n}}^{\prime}\mathbf{R}^{-1}\mathbf{U}_{\mathrm{n}} + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-\mathrm{nn}}\right)^{-1}\mathbf{U}_{\mathrm{n}}^{\prime}\mathbf{R}^{-1}\mathbf{y}^{*},\ \left(\mathbf{U}_{\mathrm{n}}^{\prime}\mathbf{R}^{-1}\mathbf{U}_{\mathrm{n}} + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-\mathrm{nn}}\right)^{-1}\right]$$

$$p(\mathbf{G}_0 \mid .) \sim IW[\nu_G + N_{\mathrm{n}}, (\mathbf{S}_G + \mathbf{V}_G)]$$

$$p(\mathbf{R}_0 \mid .) \sim IW[\nu_R + n, (\mathbf{S}_R + \mathbf{V}_R)]$$

where $\mathbf{y}^*$, S_{B_S} and $\mathbf{S}_R$ are as defined before, $\mathbf{W}_j^* = \begin{bmatrix} \mathbf{w}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_j \end{bmatrix}$, N_{n} is the number of non-genotyped individuals, and $\mathbf{S}_G = \begin{bmatrix} \boldsymbol{\epsilon}_{\mathrm{L}}^{\prime}\mathbf{A}^{-\mathrm{nn}}\boldsymbol{\epsilon}_{\mathrm{L}} & \boldsymbol{\epsilon}_{\mathrm{L}}^{\prime}\mathbf{A}^{-\mathrm{nn}}\boldsymbol{\epsilon}_{\mathrm{H}} \\ \boldsymbol{\epsilon}_{\mathrm{H}}^{\prime}\mathbf{A}^{-\mathrm{nn}}\boldsymbol{\epsilon}_{\mathrm{L}} & \boldsymbol{\epsilon}_{\mathrm{H}}^{\prime}\mathbf{A}^{-\mathrm{nn}}\boldsymbol{\epsilon}_{\mathrm{H}} \end{bmatrix}$.

Statistical analysis

Single- and multi-trait models of BayesN0 and single-step BayesN0 (ssBayesN0) were fitted with varying region sizes (one SNP, 100 SNPs, a whole chromosome and the whole genome). The parameters of the priors for SNP, residual and genetic (co) variance matrices in the multi-trait models were $\mathbf{V}_B = (\nu_B - 2 - 1)\widetilde{\mathbf{B}}$ where $\widetilde{\mathbf{B}} = \frac{\widetilde{\mathbf{G}}_0}{\sum 2p_j(1-p_j)}$, $\mathbf{V}_R = (\nu_R - 2 - 1)\widetilde{\mathbf{R}}_0$, and $\mathbf{V}_G = (\nu_G - 2 - 1)\widetilde{\mathbf{G}}_0$, which were derived from the mean of an inverse Wishart distributed random variable, and $\nu_B = \nu_R = \nu_G = 5$. It is worth noting that inverse Wishart distribution imply a scaled inverse chi-square distribution for each variance with specific parameters (Wang et al. 2018). That is, e.g., $\mathbf{B}_{s_{11}} = \sigma^2_{\alpha_{\mathrm{L},s}} \sim \chi^{-2}\left(4, \frac{\tilde{\sigma}^2_{\alpha_{\mathrm{L},s}}}{2}\right)$, where $\tilde{\sigma}^2_{\alpha_{\mathrm{L},s}}$ is the first diagonal element in $\widetilde{\mathbf{B}}$.

Single-trait BayesN0 and ssBayesN0 models were special cases of their multi-trait counterparts, for which the multivariate normal priors for SNP effects, model residuals and imputation residuals were replaced with univariate normal priors $\left(\text{e.g.}, \alpha_{\mathrm{L},sj} \sim N\left(0, \sigma^2_{\alpha_{\mathrm{L},s}}\right)\right)$, and inverse Wishart priors for the (co)variance components were replaced with scaled inverted chi-square priors $\left(\text{e.g.}, \sigma^2_{\alpha_{\mathrm{L},s}} \sim \chi^{-2}\left(\mathrm{df}, S^2_{\mathrm{L}}\right)\right)$, for conjugacy. Parameters for these scaled inverted chi-square prior distributions for SNP, residual and genetic variances were $\mathrm{df} = 4$ and a scale parameter, derived from the expected value of a scaled inverse chi-square distributed random variable $\left(\text{e.g.}, S^2_{\mathrm{L}} = \frac{\tilde{\sigma}^2_{\alpha_{\mathrm{L},s}}(\mathrm{df}-2)}{\mathrm{df}}, \text{where } \tilde{\sigma}^2_{\alpha_{\mathrm{L},s}} = \frac{\tilde{\sigma}^2_{g_{\mathrm{L}}}}{\sum 2p_j\left(1-p_j\right)}\right)$ (Habier et al. 2010a). That is, e.g., $\sigma^2_{\alpha_{\mathrm{L},s}} \sim \chi^{-2}\left(4, \frac{\tilde{\sigma}^2_{\alpha_{\mathrm{L},s}}}{2}\right)$. Hence, not only the mean, but also the distribution of priors for the variances were consistent between the single- and multi-trait analysis, with only difference being the value of variance components used. The matrices of $\widetilde{\mathbf{G}}_0$ and $\widetilde{\mathbf{R}}_0$ used in priors for multi-trait analysis, and genetic $\left(\tilde{\sigma}^2_g\right)$ and residual variances $\left(\tilde{\sigma}^2_e\right)$ used in priors for single-trait analysis, were the estimates obtained by fitting single or multi-trait Ridge-Regression models at SNP level, respectively, using the JWAS (Cheng et al. 2018a) package in Julia (Bezanson et al. 2017).

Markov-chain Monte Carlo (MCMC) algorithm with Gibbs sampling method was used to obtain samples of each parameter from its full conditional posterior distribution. Chain length for the analyses using BayesN0 and ssBayesN0 consisted of 50,000 or 70,000 cycles, of which the first 30,000 or 50,000 cycles were discarded as burn-in, respectively. Convergence was tested by comparing results for the two chain lengths (50,000 vs. 70,000) on a random subset of the replicates and region sizes (Zeng et al. 2018). Every tenth sample of the post burn-in cycles were stored for posterior analysis, yielding 2,000 posterior samples. Mean value of the posterior samples was used as the estimate of each parameter. The change in accuracy of prediction was negligible for 70,000 compared to 50,000 cycles of Markov chain, and therefore, the results from the chain length of 50,000 were presented.

For single- and multi-trait ssSNPBLUP models, the genetic and residual (co)variances and SNP (co)variances were obtained as the mean values of the posterior samples from BayesN0 or ssBayesN0. The genetic (co)variances required in mixed model equations for $\hat{\epsilon}$ were computed as the mean of the (co)variances of the breeding values at each MCMC cycle for BayesN0, or directly as the mean of genetic (co)variances for ssBayesN0. Hereafter, analysis using the variance components from BayesN0 and ssBayesN0 will be referred to as ssSNPB1 and ssSNPB2, respectively. The ssSNPB1 and ssSNPB2 models were solved with the Conjugate Gradients method with diagonal preconditioning using the IterativeSolvers package in Julia, and convergence tolerance was chosen to be 10^{-12}. All analyses were performed using self-written scripts in Julia.

The predicted breeding values of animals using multi-trait BayesN0 were obtained from

$$\widehat{\mathbf{g}}_t = \mathbf{M}\widehat{\boldsymbol{\alpha}}_t, \quad t = \mathrm{L}, \mathrm{H}.$$

The predicted breeding values of animals using single-step models, ssBayesN0, ssSNPB1 and ssSNPB2, were obtained from:

$$\widehat{\mathbf{g}}_t = \begin{bmatrix} -\mathbf{A}_{\mathrm{ng}}\mathbf{A}_{\mathrm{gg}}^{-1}\mathbf{1} \\ -\mathbf{1} \end{bmatrix}\hat{\mu}_{\mathrm{g},t} + \begin{bmatrix} \widehat{\mathbf{M}}_{\mathrm{n}} \\ \mathbf{M}_{\mathrm{g}} \end{bmatrix}\widehat{\boldsymbol{\alpha}}_t + \begin{bmatrix} \mathbf{Z}_{\mathrm{n}} \\ \mathbf{0} \end{bmatrix}\widehat{\boldsymbol{\epsilon}}_t, \quad t = \mathrm{L}, \mathrm{H} \tag{4}$$

Prediction accuracy was assessed as the correlation between true and predicted breeding values of validation individuals. The bias of prediction was assessed based on the slope of the regression of true breeding values on the estimated breeding values of validation individuals. Accuracy for single- and multi-trait models with different region sizes were compared for each trait, and each model separately. Prediction accuracy for all methods was compared for each trait and at each scenario of region size. All comparisons were performed separately for genotyped and non-genotyped individuals using a two-sided paired *t*-tests, for which accuracies were paired across each replicate for the same validation population. A Bonferroni correction was used to control the Type 1 error rate of 0.05, caused by multiple comparisons.

Results

Bayesian whole genome regression (BayesN0)

Prediction accuracies from single- and multi-trait BayesN0 models are given in Tables 2 and 4 for genotyped individuals in validation population, at varying sizes of genome region. Grouping 100 adjacent SNPs generally provided the highest accuracies for both single- and multi-trait models, with some exceptions in scenario N5. Accuracies for different region sizes were generally ranked as 100 SNPs > 1 SNP > 1 Chr > WG in scenario G9. When a multi-trait model was used in scenario G9, prediction accuracy for the region size of 100 SNPs were about 4 and 12 percentage points higher for low heritability trait (L), and about 3 and 8 percentage points higher for high heritability trait (H), compared to those for region sizes of one SNP (BayesA) and whole genome (GBLUP), respectively. Using multi-trait BayesN0 with a region size of 100 SNPs resulted in higher accuracies than corresponding single-trait BayesN0 for both traits, though not always significant. Bias for predicting breeding values of genotyped individuals is shown in Supplementary Tables S1 and S3. Regression coefficients were generally closer to 1 for trait H in both scenarios. They were higher than 1 particularly for single-trait analysis of trait L in scenario G9 and single- and multi-trait analysis of trait L in scenario N5.

Single-step genomic prediction

Prediction accuracies from single- and multi-trait analysis are given in Tables 2–5. Similar to BayesN0, accuracies for different region sizes were generally ranked as 100 SNPs > 1 SNP > 1 Chr > WG in scenario G9. For single-trait analysis of trait L, accuracies from the region size of 1 SNP and/or WG were similar to, or even slightly higher than, that of region size of 100 SNPs in scenario N5. Using ssSNPB1 improved accuracies for genotyped individuals compared to using BayesN0, for both single- and multi-trait analysis. Accuracies from ssBayesN0 were generally similar to or somewhat higher than those from ssSNPB1, particularly in scenario G9. Using single-step SNPBLUP with (co)variances obtained from ssBayesN0, i.e., ssSNPB2, yielded similar accuracies to the corresponding ssBayesN0 model. Accuracies from ssSNPB2 were similar to, though sometimes slightly higher in scenario G9, those from ssSNPB1 for non-genotyped animals. For non-genotyped animals, taking 100 adjacent SNPs as a genome region provided similar to or slightly higher accuracies

Table 2 Accuracies for genotyped individuals using single- and multi-trait models in scenario G9

Trait[1]	Region size[2]	Single-trait[3]				Multi-trait			
		BayesN0	ssSNPB1	ssBayesN0	ssSNPB2	BayesN0	ssSNPB1	ssBayesN0	ssSNPB2
L[4]	1 SNP	${}_{ab}0.349^{e}$	${}_{ab}0.452^{cd}$	${}_{ab}0.470^{d}$	${}_{ab}0.478^{c}$	${}_{b}0.437^{d}$	${}_{b}0.536^{b}$	${}_{b}0.554^{b}$	${}_{a}0.574^{a}$
	100 SNPs	${}_{a}0.365^{d}$	${}_{a}0.460^{c}$	${}_{a}0.478^{c}$	${}_{a}0.479^{c}$	${}_{a}0.481^{c}$	${}_{a}0.559^{b}$	${}_{a}0.590^{a}$	${}_{a}0.590^{a}$
	1 Chr	${}_{b}0.335^{c}$	${}_{b}0.434^{b}$	${}_{bc}0.444^{b}$	${}_{bc}0.445^{b}$	${}_{b}0.402^{b}$	${}_{c}0.493^{a}$	${}_{c}0.497^{a}$	${}_{b}0.499^{a}$
	WG	${}_{b}0.335^{d}$	${}_{b}0.433^{b}$	${}_{c}0.433^{bc}$	${}_{c}0.433^{b}$	${}_{c}0.362^{cd}$	${}_{d}0.461^{ab}$	${}_{d}0.472^{a}$	${}_{c}0.473^{a}$
H	1 SNP	${}_{b}0.587^{g}$	${}_{b}0.683^{e}$	${}_{b}0.689^{de}$	${}_{b}0.700^{bc}$	${}_{b}0.593^{f}$	${}_{b}0.688^{cd}$	${}_{b}0.699^{b}$	${}_{a}0.712^{a}$
	100 SNPs	${}_{a}0.611^{f}$	${}_{a}0.698^{d}$	${}_{a}0.716^{b}$	${}_{a}0.716^{b}$	${}_{a}0.622^{e}$	${}_{a}0.707^{c}$	${}_{a}0.725^{a}$	${}_{a}0.725^{a}$
	1 Chr	${}_{c}0.552^{e}$	${}_{c}0.650^{bd}$	${}_{c}0.651^{cd}$	${}_{c}0.651^{abcd}$	${}_{c}0.558^{e}$	${}_{c}0.655^{ac}$	${}_{c}0.657^{ab}$	${}_{b}0.657^{ab}$
	WG	${}_{d}0.538^{d}$	${}_{c}0.642^{c}$	${}_{c}0.642^{bc}$	${}_{c}0.643^{bc}$	${}_{d}0.543^{d}$	${}_{d}0.644^{abc}$	${}_{d}0.646^{ab}$	${}_{c}0.646^{a}$

[1] L and H: low (0.1) and high (0.4) heritability traits, respectively

[2] Chr chromosome, WG whole genome

[3] ssSNPB1 and ssSNPB2: Single-step SNPBLUP, for which the variance components were obtained from BayesN0 and ssBayesN0, respectively

[4] Different alphabets mean significantly different values at a Type 1 error rate of 0.05 with Bonferroni correction. Subscripts and superscripts stand for comparisons within column and row, respectively, for each trait

Table 3 Accuracies for non-genotyped individuals using single- and multi-trait models in scenario G9

Trait[1]	Region size[2]	Single-trait[3]			Multi-trait		
		ssSNPB1	ssBayesN0	ssSNPB2	ssSNPB1	ssBayesN0	ssSNPB2
L[4]	1 SNP	${}_{ab}0.351^{c}$	${}_{ab}0.357^{c}$	${}_{ab}0.361^{c}$	${}_{b}0.402^{b}$	${}_{b}0.406^{b}$	${}_{a}0.418^{a}$
	100 SNPs	${}_{a}0.355^{c}$	${}_{a}0.363^{c}$	${}_{a}0.363^{c}$	${}_{a}0.412^{b}$	${}_{a}0.426^{a}$	${}_{a}0.427^{a}$
	1 Chr	${}_{b}0.340^{b}$	${}_{b}0.342^{b}$	${}_{bc}0.342^{b}$	${}_{c}0.378^{a}$	${}_{c}0.377^{a}$	${}_{b}0.378^{a}$
	WG	${}_{b}0.337^{c}$	${}_{b}0.337^{c}$	${}_{c}0.336^{c}$	${}_{d}0.361^{abc}$	${}_{d}0.365^{b}$	${}_{c}0.366^{a}$
H	1 SNP	${}_{a}0.526^{cd}$	${}_{a}0.528^{d}$	${}_{a}0.531^{bc}$	${}_{a}0.529^{bcd}$	${}_{a}0.533^{b}$	${}_{a}0.537^{a}$
	100 SNPs	${}_{a}0.530^{c}$	${}_{a}0.535^{b}$	${}_{a}0.535^{b}$	${}_{a}0.534^{b}$	${}_{a}0.539^{a}$	${}_{a}0.539^{a}$
	1 Chr	${}_{b}0.513^{abc}$	${}_{b}0.513^{c}$	${}_{b}0.513^{bc}$	${}_{b}0.516^{abc}$	${}_{b}0.517^{ab}$	${}_{b}0.517^{a}$
	WG	${}_{b}0.511^{bc}$	${}_{b}0.511^{c}$	${}_{b}0.511^{bc}$	${}_{b}0.513^{abc}$	${}_{b}0.514^{ab}$	${}_{c}0.514^{a}$

[1] L and H: low (0.1) and high (0.4) heritability traits, respectively

[2] Chr chromosome, WG whole genome

[3] ssSNPB1 and ssSNPB2: Single-step SNPBLUP, for which the variance components were obtained from BayesN0 and ssBayesN0, respectively

[4] Different alphabets mean significantly different values at a Type 1 error rate of 0.05 with Bonferroni correction. Subscripts and superscripts stand for comparisons within column and row, respectively, for each trait

Table 4 Accuracies for genotyped individuals using single- and multi-trait models in scenario N5

Trait[1]	Region size[2]	Single-trait[3]				Multi-trait			
		BayesN0	ssSNPB1	ssBayesN0	ssSNPB2	BayesN0	ssSNPB1	ssBayesN0	ssSNPB2
L[4]	1 SNP	${}_{a}0.314^{d}$	${}_{a}0.432^{b}$	${}_{a}0.432^{b}$	${}_{a}0.433^{b}$	${}_{a}0.362^{c}$	${}_{a}0.470^{a}$	${}_{a}0.469^{a}$	${}_{a}0.470^{a}$
	100 SNPs	${}_{a}0.313^{e}$	${}_{a}0.428^{bc}$	${}_{a}0.434^{b}$	${}_{ab}0.429^{c}$	${}_{a}0.367^{d}$	${}_{ab}0.468^{a}$	${}_{a}0.475^{a}$	${}_{a}0.474^{a}$
	1 Chr	${}_{a}0.309^{d}$	${}_{b}0.419^{c}$	${}_{b}0.420^{c}$	${}_{b}0.419^{c}$	${}_{b}0.341^{d}$	${}_{c}0.447^{abc}$	${}_{b}0.447^{b}$	${}_{b}0.450^{a}$
	WG	${}_{a}0.314^{c}$	${}_{a}0.431^{ab}$	${}_{ab}0.430^{b}$	${}_{a}0.432^{a}$	${}_{b}0.342^{c}$	${}_{bc}0.447^{ab}$	${}_{ab}0.459^{ab}$	${}_{ab}0.460^{ab}$
H	1 SNP	${}_{b}0.545^{d}$	${}_{a}0.651^{c}$	${}_{b}0.654^{bc}$	${}_{a}0.658^{ab}$	${}_{b}0.548^{d}$	${}_{a}0.654^{bc}$	${}_{b}0.657^{b}$	${}_{a}0.662^{a}$
	100 SNPs	${}_{a}0.554^{d}$	${}_{a}0.654^{c}$	${}_{a}0.663^{a}$	${}_{a}0.662^{ab}$	${}_{a}0.559^{d}$	${}_{a}0.657^{bc}$	${}_{a}0.666^{a}$	${}_{a}0.665^{a}$
	1 Chr	${}_{c}0.537^{b}$	${}_{b}0.642^{a}$	${}_{c}0.645^{a}$	${}_{b}0.645^{a}$	${}_{c}0.540^{b}$	${}_{b}0.644^{a}$	${}_{c}0.646^{a}$	${}_{b}0.647^{a}$
	WG	${}_{c}0.537^{b}$	${}_{b}0.644^{a}$	${}_{c}0.645^{a}$	${}_{b}0.646^{a}$	${}_{c}0.539^{b}$	${}_{b}0.645^{a}$	${}_{c}0.648^{a}$	${}_{b}0.648^{a}$

[1] L and H: low (0.1) and high (0.4) heritability traits, respectively

[2] Chr chromosome, WG whole genome

[3] ssSNPB1 and ssSNPB2: Single-step SNPBLUP, for which the variance components were obtained from BayesN0 and ssBayesN0, respectively

[4] Different alphabets mean significantly different values at a Type 1 error rate of 0.05 with Bonferroni correction. Subscripts and superscripts stand for comparisons within column and row, respectively, for each trait

Table 5 Accuracies for non-genotyped individuals using single- and multi-trait models in scenario N5

Trait[1]	Region size[2]	Single-trait[3]			Multi-trait		
		ssSNPB1	ssBayesN0	ssSNPB2	ssSNPB1	ssBayesN0	ssSNPB2
L[4]	1 SNP	${}_{a}0.328^{c}$	${}_{a}0.325^{c}$	${}_{a}0.326^{c}$	${}_{a}0.357^{a}$	${}_{a}0.353^{b}$	${}_{a}0.355^{ab}$
	100 SNPs	${}_{a}0.327^{b}$	${}_{a}0.327^{b}$	${}_{a}0.326^{b}$	${}_{a}0.357^{a}$	${}_{a}0.355^{a}$	${}_{a}0.355^{a}$
	1 Chr	${}_{a}0.324^{c}$	${}_{a}0.325^{c}$	${}_{a}0.324^{c}$	${}_{ab}0.349^{ab}$	${}_{a}0.347^{b}$	${}_{a}0.350^{a}$
	WG	${}_{a}0.327^{b}$	${}_{a}0.325^{b}$	${}_{a}0.327^{ab}$	${}_{b}0.343^{ab}$	${}_{a}0.350^{b}$	${}_{a}0.352^{a}$
H	1 SNP	${}_{a}0.506^{d}$	${}_{b}0.507^{cd}$	${}_{a}0.510^{b}$	${}_{a}0.508^{bc}$	${}_{ab}0.509^{b}$	${}_{a}0.512^{a}$
	100 SNPs	${}_{a}0.507^{c}$	${}_{a}0.511^{ab}$	${}_{a}0.511^{ab}$	${}_{a}0.509^{bc}$	${}_{a}0.512^{ab}$	${}_{a}0.512^{a}$
	1 Chr	${}_{ab}0.503^{b}$	${}_{bc}0.504^{ab}$	${}_{b}0.504^{ab}$	${}_{ab}0.505^{a}$	${}_{bc}0.505^{ab}$	${}_{b}0.506^{ab}$
	WG	${}_{b}0.503^{b}$	${}_{c}0.503^{b}$	${}_{b}0.503^{b}$	${}_{b}0.504^{ab}$	${}_{c}0.506^{a}$	${}_{b}0.506^{a}$

[1] L and H: low (0.1) and high (0.4) heritability traits, respectively

[2] Chr chromosome, WG whole genome

[3] ssSNPB1 and ssSNPB2: Single-step SNPBLUP, for which the variance components were obtained from BayesN0 and ssBayesN0, respectively

[4] Different alphabets mean significantly different values at a Type 1 error rate of 0.05 with Bonferroni correction. Subscripts and superscripts stand for comparisons within column and row, respectively, for each trait

than taking one SNP as a genome region, but higher accuracies than taking whole genome as a genome region, in scenario G9. For scenario N5, on the other hand, all region sizes generally lead to similar accuracies for non-genotyped animals. Regression coefficients were generally closer to 1 for trait H, but higher than 1 for trait L in scenario N5 (Supplementary Tables S1–S4).

Discussion

Single- vs. multi-trait genomic prediction

Multi-trait analysis generally led to higher accuracies than their single-trait counterparts for trait L ($h^2 = 0.1$), and similar to or higher accuracies than their single-trait counterparts for trait H ($h^2 = 0.4$) (Tables 2–5). This was expected because the gain of accuracy from multi-trait over single-trait genomic prediction is more profound for low heritability traits that are genetically correlated with a high heritability trait (Jia and Jannink 2012; Guo et al. 2014). Hayashi and Iwata (2013) compared accuracies from single- and multi-trait analysis for traits with a genetic correlation of 0.7, and reported that accuracy for a low heritability trait ($h^2 = 0.1$) was improved with multi-trait analysis, while accuracy for a high heritability ($h^2 = 0.8$) trait remained unchanged. For a low heritability ($h^2 = 0.05$) trait, which had incomplete data, Guo et al. (2014) showed that accuracy of genomic prediction was improved when a genetically correlated ($r_g = 0.5$) trait with high heritability

($h^2 = 0.3$) was available. Cheng et al. (2018b) reported that the mean of the posterior probability that a marker has a null effect was higher (0.97 vs. 0.74) in multi-trait analysis (BayesCΠ) compared to single-trait analysis (BayesCπ) for gall volume ($h^2 = 0.12$), when the correlated trait was presence (or absence) of rust ($h^2 = 0.21$), in Loblolly Pine (*Pinus taeda* L.) (Resende et al. 2012).

Beside heritability, another factor influencing accuracy is the absolute difference between genetic and residual correlations (Schaeffer 1984; Thompson and Meyer 1986). In this study, the simulated residual correlation was null and the genetic correlation was moderate (0.45), though the estimates of those correlations varied around the simulated true values. Averaged over the replicates, genetic correlations were generally overestimated, whereas the residual correlations were nearly zero and varied only after second decimal, in both scenarios and for all region sizes. Genetic correlations were 0.47 and 0.45 from BayesN0 with the region size of 100 SNPs, and 0.56 and 0.51 from GBLUP (BayesN0 with whole genome as one region), for scenarios G9 and N5, respectively (results not given elsewhere). Those were 0.49 and 0.47 for ssBayesN0 with the region size of 100 SNPs, and 0.54 and 0.48 for ssGBLUP (ssBayesN0 with whole genome as one region), for scenarios G9 and N5, respectively (results not given elsewhere). These small deviations of genetic correlations from their true values are expected to have little influence in variance of prediction error (PEV), and multi-trait models can increase the precision of breeding value estimates by reducing PEV compared to single-trait models (Schaeffer 1984). The PEV was additionally computed for BayesN0 and ssBayesN0, from the variance of posterior samples for breeding values of genotyped individuals in validation population. Averaged over region sizes, the mean reduction in PEV from multi-trait BayesN0 were about 2.5% for trait L and 0.5% for trait H, and 5% for trait L and 0.5% for trait H, in scenarios G9 and N5, respectively (results not given elsewhere). The mean reduction in PEV from multi-trait ssBayesN0 were about 9% for trait L and 0.9% for trait H, and 6% for trait L and 0.8% for trait H, in scenarios G9 and N5, respectively (results not given elsewhere). Bias for single-trait analysis was relatively high for trait L particularly in scenario G9 (Supplementary Tables S1–S4), however, it was generally reduced by using multi-trait models.

In multi-trait genomic prediction, correlation structures between the traits is central to gaining advantage in prediction accuracy over single-trait predictions (Gebreyesus et al. 2017). Our results showed that the improvement from multi-trait analysis over single-trait analysis were dependent on whether the genetic makeup of the (co)variance structure of the studied traits (Tables 2–5) were accounted for, and this will be discussed in detail in the later sections.

Accounting for heterogeneous (co)variances across the genome using BayesN0

Multi-trait genomic prediction rely on the genetic association between the traits through the genetic variances and covariances, which may vary across the genome. A few genome regions may explain a substantial proportion of the covariance, whereas others account for nearly no covariance between the traits (Sørensen et al. 2012). Moreover, covariances between particular traits may be positive for some regions and negative for others, while the overall genetic correlations are low/high (Li et al. 2017; Gebreyesus et al. 2017). This study investigated the affect of assigning priors to genome regions, which were defined as fixed number of SNPs (one SNP, 100 SNPs, one chromosome or whole genome), on accuracy in multi-trait genomic prediction.

Genomic prediction rests on the LD between QTL and SNPs (Meuwissen et al. 2001). Although the simulation settings in this study resulted in correlations of QTL effects that fall into different categories, it may be of a general question where does the heterogenity of (co)variances over the genome come from, or what does it refer to. It can be shown that the best linear predictor of SNP effects is $\boldsymbol{\alpha}_t = \mathbf{V}_{\mathrm{M}}^{-1}\mathbf{V}_{\mathrm{MQ}}\boldsymbol{\gamma}_t$ ($t = \mathrm{L}, \mathrm{H}$), where $\boldsymbol{\gamma}_t$ is the vector of QTL effects, $\mathbf{V}_{\mathrm{M}}$ is the (co)variance matrix of SNP genotypes, and $\mathbf{V}_{\mathrm{MQ}}$ is the covariance matrix of SNP and QTL genotypes (de los Campos et al. 2015). Note that for a QTL that affect only L (or H), corresponding row of $\boldsymbol{\gamma}_{\mathrm{H}}$ (or $\boldsymbol{\gamma}_{\mathrm{L}}$) is zero. Under some assumptions, (co)variance of the SNP effects are proportional to $\mathbf{V}_{\mathrm{M}_s\mathrm{Q}_s}\mathbf{V}'_{\mathrm{M}_s\mathrm{Q}_s}$, for genome region s ($s = 1, \ldots, S$). Because recombination rates vary over the genome, and SNPs are typically in imperfect LD with QTL, each $V_{M_s}Q_s$ may be different (Wang et al. 2013), resulting in genome having a different (co)variance pattern at the SNP level than that of at the QTL level (de los Campos et al. 2015).

Multi-trait BayesA (BayesN0 with region size of one SNP) was able to account for the heterogeneous correlation structure across the genome to some extent, compared to multi-trait GBLUP (BayesN0 with whole genome as one region), which assumes a constant correlation across the genome (Tables 2 and 4). Accuracies were further improved when a group of 100 SNPs were allowed to have a common (co)variance. It should be noted that the choice of region sizes was arbitrary, and therefore, the region size of 100 SNPs may not be optimal. Alternatively, regions can be achieved by grouping SNPs based on fixed length of genomic region or LD information. Because the extent of LD is highly variable in different populations (Wang et al. 2013), and varies with respect to SNP density (Goddard and Hayes 2009), the decision of optimal region size is crucial to obtain highest accuracy of genomic prediction (Gebreyesus et al. 2017).

Simulation studies have shown that Bayesian whole genome regression models, which allow variances of SNP effects differing among loci or genome regions, perform better than GBLUP model (Meuwissen et al. 2001; Lund et al. 2009; Karaman et al. 2018). In real-data applications, the accuracy of genomic prediction using the Bayesian whole genome regression models led to similar to or higher accuracies than methods assuming a constant variance structure (e.g., GBLUP) across the genome (Hayes et al. 2009; Habier et al. 2010b; Su et al. 2012a). The benefit from Bayesian whole genome regression models was larger for traits with simple genetic architectures (Coster et al. 2010; Daetwyler et al. 2010b; Clark et al. 2011; Karaman et al. 2018). Examples for such traits can be milk protein composition traits, in which a substantial proportion of the variance is explained by a few QTL (Heck et al. 2009; Schopen et al. 2011). Gebreyesus et al. (2017) reported that BaysesAS model resulted in higher prediction reliabilities than GBLUP for milk protein composition traits, when 100 SNPs were assumed to have a common (co)variance, based on a data set from 50 K SNP panel in Danish Holstein cattle.

For prediction of traits with large effect QTL, the GBLUP model, in which a selection of SNPs, i.e., SNPs identified in earlier genome-wide association studies (GWAS) or identified via GWAS using the current data (*de novo* GWAS, Spindel et al. (2016)), are considered as fixed effects, can provide accuracies as high as those from Bayesian whole genome prediction methods (Spindel et al. 2016; Lopes et al. 2017). Although the approach is relatively straightforward, it either requires a priori information about the SNPs for the traits of interest, or running a GWAS prior to genomic prediction. Depending on the choice of statistical method, the definition of the QTL region and the significance threshold, different sets of SNPs can be achieved even with the same data, and QTL regions that explain a substantial proportion of the variance may also not always be identified for all traits (Goddard et al. 2016; Lopes et al. 2017).

By applying BayesN0, one has the possibility of putting emphasis on genome regions with large effect, without requiring any prior knowledge on the QTL region affecting the trait(s), or without running a *de novo* GWAS (Lopes et al. 2017). For practical application in breeding programs, we think this is an advantage over GBLUP, in which "some" SNPs are considered as fixed effects. Our results for scenario N5 imply that the advantage of grouping SNPs in BayesN0 over GBLUP is not limited only to traits with a few QTL with large effect and many with small effects (scenario G9). In scenario N5, BayesN0 with 100 SNPs region size led to a similar accuracy to that from GBLUP in single-trait analysis of trait L, but to a higher accuracy than GBLUP in multi-trait analysis of trait L. Since using a multi-trait model may be beneficial for traits by increasing the amount of information, it can be argued that the accuracies from single-trait analysis of trait L would also differ among the region sizes, for the intermediate sizes of data (Karaman et al. 2016). For asymptotically large sizes of data, on the other hand, there might be little or no benefit of using more sophisticated methods compared to GBLUP (Karaman et al. 2016; Cheng et al. 2018b).

In a simulation study for single-trait genomic prediction, Zeng et al. (2018) showed that BayesN was superior to BayesB when the QTL had relatively low MAF, for a panel consisting of 50 K SNPs. It is, however, unclear if this was due to selection of regions at each cycle of MCMC, or due to reliable estimation of SNP variances by assuming common variance to SNPs in each region, rather than assuming a variance specific to each SNP. In that study, fitting ten SNPs per region also provided higher accuracies of prediction than fitting two SNPs per region. Hess et al. (2017) further allowed SNPs within a region to have different variances, in a study using 50 K SNP panel of an admixed cattle population in New Zealand. There was no advantage of BayesN over BayesB, for milk fat yield, live-weight and somatic cell score. They also showed that fitting all SNPs in a region resulted in slightly higher accuracies than fitting only two SNPs per region.

Accounting for heterogeneous (co)variances across the genome using single-step Bayesian regression

Implementation of our novel Bayesian multi-trait model (ssBayesN0) using the methodology of Fernando et al. (2014) yielded accuracies for genotyped individuals in the range of 0.47–0.59 and 0.45–0.48 for trait L, and 0.65–0.73 and 0.65–0.67 for trait H, in scenarios G9 and N5, respectively (Tables 2 and 4). In a single-step analysis using Bayesian regression (Fernando et al. 2014), taking one SNP as a genome region is equivalent to single-step BayesA (ssBayesA) and taking whole genome as one region is equivalent to single-step GBLUP (ssGBLUP). Our results indicate that ssBayesA can lead to higher accuracies than ssGBLUP in a multi-trait analysis, by exploiting the heterogeneous (co)variance structure across the genome. However, similar to the regular BayesA (Meuwissen et al. 2001), the information in the data that is utilized by ssBayesA is limited, due to its strong dependency on the prior for (co)variance of SNP effects (Gianola et al. 2009). This dependency on the prior was overcome to some extent by assuming a common (co)variance for 100 adjacent SNPs using ssBayesN0, which generally led to higher accuracies than ssBayesA and ssGBLUP for both trait L and H (Tables 2 and 4). Similarly, prediction accuracy for non-genotyped individuals were increased about 6 and 0.5 percentage points for trait L, and about 2.5 and 0.6 percentage points for trait H, for scenarios G9 and N5, respectively, when region size was changed from whole genome to 100 SNPs in multi-trait analyses (Tables 3 and 5).

For genotyped individuals, using multi-trait ssBayesN0 led to higher accuracies than using multi-trait BayesN0 (Tables 2 and 4). As other Bayesian whole genome regression models, BayesN0 can only use the phenotypes of genotyped animals. The ssBayesN0, on the other hand, simultaneously uses the phenotypes of genotyped animals (1000) and non-genotyped animals (3000, the genotypes were imputed) (Table 1) to estimate the SNP effects in the MCMC procedure, while accounting for the error in imputation for non-genotyped individuals with phenotypes. This enhances the data size used in estimation of SNP effects, which has a key role to obtain reliable prediction of breeding values (Daetwyler et al. 2008; Goddard 2009; Karaman et al. 2016; Cheng et al. 2018b).

Practical implementation of single-step models using previously estimated (co)variance components

In this study, the estimates of the (co)variances were obtained from BayesN0 or ssBayesN0. The former led to ssSNPB1 model for which the (co)variance components were obtained using only the information of genotyped individuals, while the latter led to ssSNPB2 model for which the (co)variance components were obtained using the information of genotyped and non-genotyped individuals. In practice, the (co)variance components can be estimated less frequently compared to routine genomic evaluations without harming the prediction accuracies (Su et al. 2014).

For genotyped individuals, ssSNPB1 model yielded higher accuracies than BayesN0, at all region sizes in multi-trait analysis (Tables 2 and 4). This was due to more accurate estimation of SNP effects by the use of phenotypes of non-genotyped individuals. The ssBayesN0 and ssSNPB2, where the (co)variance components from ssBayesN0 were used, generally yielded similar accuracies. This was not surprising, because similar to BayesC0 and SNPBLUP being equivalent models, ssBayesN0 and ssSNPB2, are also equivalent. The ssSNPB2 generally led to higher accuracies than ssSNPB1 in scenario G9, and similar to or slightly higher accuracies than ssSNPB1 in scenario N5, in multi-trait analyses.

Accuracies for non-genotyped animals were generally similar among the models, i.e., ssSNPB1, ssBayesN0 and ssSNPB2, in multi-trait analysis. Analysis using the model of Fernando et al. (2014) starts with an explicit imputation of markers for non-genotyped individuals, using pedigree information and genotypes of genotyped relatives. Then, marker effects and imputation residuals (ϵ) accounting for the part of breeding values, which cannot be modeled by imputed markers, are estimated (Gao et al. 2018). Imputation residual is added to marker-based breeding value (sum of individual SNP effects) of non-genotyped individuals to obtain their total breeding values (Fernando et al. 2014). The breeding value of genotyped individuals, on the other hand, is composed only of sum of individual SNP effects. Hence, a change in the accuracy of SNP effect estimates has less impact on the accuracy of breeding value estimates for non-genotyped individuals than for genotyped individuals (Zhou et al. 2018).

One way to account for heterogeneous (co)variance structure in single-step genomic prediction could be to construct weighted **G** matrices (Zhang et al. 2010), and in turn their weighted **H** matrix counterparts (Fragomeni et al. 2017) for ssGBLUP. In an earlier study (Karaman et al. 2018), we have shown that weighted multi-trait GBLUP can reach accuracies similar to that of the Bayesian whole genome regression model which was used to derive weights. This was expected, because those “weighted” relationship matrices are indeed implicit to Bayesian whole genome regression methods (Fernando and Gianola 2018; Karaman et al. 2018). A drawback of the approach using weighted relationship matrices is that it requires the computation of a number of relationship matrices which increase with the number of traits in a multi-trait model, and not only the computing time but also the storage of such **H** matrices might be impractical for genomic prediction using weighted ssGBLUP in routine evaluations. Moreover, compared to ssGBLUP, the equations needed to be solved for ssSNPBLUP does not grow with the number of genotyped individuals, and the inverse of the combined relationship matrix, **H**, is not needed (Fernando et al. 2014).

We did not focus on the computational (dis)advantages of ssSNPBLUP, nor its convergency properties. Both ssGBLUP and ssSNPBLUP, though, are known to have some computational challenges (Taskinen et al. 2017). Averaged over the scenarios and region sizes, ssSNPB1 and ssSNPB2 models achieved relative convergence of 10^{-12} in 200 and 203 iterations (average of two traits) in single-trait analysis, and in 435 and 478 iterations in multi-trait analysis, respectively. It should be noted that these numbers apply only to the current data, and could vary with equivalent formulations of the models (Taskinen et al. 2017) or with a preconditioner other than diagonal used in this study.

Conclusions

In this study, a multi-trait whole genome regression model, BayesN0, was proposed. The model has its equivalent counterparts when the region size is set at one SNP (BayesA) or the whole genome (GBLUP). Our results showed that assigning priors to genome regions defined as fixed number of SNPs, e.g., 100 SNPs, may improve accuracies over BayesA and GBLUP by accounting for heterogeneous (co)variance structure across the genome efficiently. The model was also implemented in single-step (ssBayesN0) Bayesian regression approach, which unifies pedigree, phenotypes and genotypes in a single analysis. Highest prediction accuracies were obtained when 100 adjacent SNPs were assumed to have a common (co)variance in ssBayesN0. For routine genomic evaluations, it could be a good strategy to estimate (co)variance components from ssBayesN0, and then to use those estimates in genomic prediction using multi-trait single-step SNPBLUP. Such a strategy has the potential to provide reliable estimates of breeding values for both genotyped and non-genotyped individuals.

Data availability

Genotype and pedigree data can be found at https://doi.org/10.5061/dryad.v4126t4, along with a file including necessary SNP information (chromosome ID and base-pair position). The data and the methodology described previously are sufficient to reproduce the results of this study.

Acknowledgements This study was funded by the "Multi-Genomics" project from the Danish Milk Levy Fund (Aarhus, Denmark), and "Genomics in Herds" project financed by VikingGenetics (Randers, Denmark) and Nordic Cattle Evaluation (Aarhus, Denmark). The first author is grateful to Dr. Hao Cheng (Department of Animal Science, University of California Davis) for his support in using JWAS package, and making his codes generously available. The first author also acknowledges fruitful discussions with Dr. Aoxing Liu (Department of Molecular Biology and Genetics, Aarhus University).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Aguilar I, Misztal I, Johnson D, Legarra A, Tsuruta S, Lawlor T (2010) A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci 93:743–752

Bezanson J, Edelman A, Karpinski S, Shah V (2017) Julia: a fresh approach to numerical computing. SIAM Rev 59:65–98

Brøndum RF, Su G, Lund M, Bowman P, Goddard M, Hayes B (2012) Genome position specific priors for genomic prediction. BMC Genomics 13:543

Calus M, Schrooten C, Veerkamp R (2014) Genomic prediction of breeding values using previously estimated SNP variances. Genet Sel Evol 46:52

Calus MP, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol 43:26

Cheng H, Fernando R, Garrick D (2018a) JWAS: Julia implementation of whole-genome analysis software. World Congr Genet Appl Livest Prod 11:859

Cheng H, Kizilkaya K, Zeng J, Garrick D, Fernando R (2018b) Genomic prediction from multiple-trait Bayesian regression methods using mixture priors. Genetics 209:89–103

Christensen O, Lund M (2010) Genomic prediction when some animals are not genotyped. Genet Sel Evol 42:2

Clark S, Hickey J, van der Werf H (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43:18

Coster A, JW B, Calus M, van Arendonk JA, Bovenhuis H (2010) Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet Sel Evol 42:9

Daetwyler H, Hickey J, Henshall J, Dominik S, Gredler B et al. (2010a) Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Anim Prod Sci 50:1004–1010

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

Daetwyler H, Villanueva B, Woolliams J (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3:e3395

de los Campos G, Sorensen D, Gianola D (2015) Genomic heritability: what is it? PLoS Genet 11:e1005048

Dvorkin D (2012) lcmix: Layered and chained mixture models. R package version 03/r5. https://r-forge.r-project.org/R/?group_id=1092

Fernando R, Dekkers J, Garrick D (2014) A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel Evol 46:50

Fernando R, Gianola D (2018) Bayesian inference of genomic similarity among individuals from markers and phenotypes. In: Proceedings of the World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand. p 942

Fragomeni BO, Lourenco DAL, Masuda Y, Legarra A, Misztal I (2017) Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genet Sel Evol 49:59

Gao H, Koivula M, Jensen J, Strandén I, Madsen P, Pitkänen T et al. (2018) Short communication: genomic prediction using different single-step methods in the Finnish red dairy cattle population. J Dairy Sci 101:10082–10088

Gebreyesus G, Lund M, Buitenhuis B, Bovenhuis H, Poulsen N, Janss L (2017) Modeling heterogeneous (co)variances from adjacent-SNP groups improves genomic prediction for milk protein composition traits. Genet Sel Evol 49:89

Gianola D, de los Campos G, Hill W, Manfredi E, Fernando R (2009) Additive genetic variability and Bayesian alphabet. Genetics 183:347–363

Goddard M (2009) Genomic selection: prediction of accuracy and maximization of long term response. Genetica 136:245–257

Goddard M, Hayes B (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381–391

Goddard M, Kemper K, MacLeod I, Chamberlain A, Hayes B (2016) Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. Proc R Soc B 283:pii: 20160569

Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G (2014) Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet 15:30

Habier D, Fernando R, Dekkers J (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397

Habier D, Fernando RL, Kizilkaya K, Garrick DJ(2010a) Extension of the Bayesian alphabet for genomic selection BMC Bioinform 12:186

Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G(2010b) The impact of genetic relationship information on genomic breeding values in German Holstein cattle Genet Sel Evol 42:5

Hayashi T, Iwata H (2013) A bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinform 14:34

Hayes B, Bowman P, Chamberlain A, Goddard M (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443

Heck J, Schennink A, van Valenberg H, Bovenhuis H, Visker M, van Arendonk J et al. (2009) Effects of milk protein variants on the protein composition of bovine milk. J Dairy Sci 92:1192–1202

Henderson C (1984) Applications of linear models in animal breeding. University Guelph, Guelph, Ontario, Canada

Hess M, Druet T, Hess A, Garrick D (2017) Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genet Sel Evol 49:54

Jia Y, Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513–1522

Karaman E, Cheng H, Firat M, Garrick D, Fernando R (2016) An upper bound for accuracy of prediction using GBLUP. PLoS ONE 11:e0161054

Karaman E, Lund M, Anche M, Janss L, Su G (2018) Genomic prediction using multi-trait weighted GBLUP accounting for heterogeneous variances and covariances across the genome. G3-Genes Genom Genet 8:3549–3558

Kemper K, Goddard M (2012) Understanding and predicting complex traits: knowledge from cattle. Hum Mol Genet 21:45–51

Kizilkaya K, Fernando R, Garrick D (2010) Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J Anim Sci 88:544–551

Li X, Lund M, Janss L, Wang C, Ding X, Zhang Q et al. (2017) The patterns of genomic variances and covariances across genome for milk production traits between Chinese and Nordic Holstein populations. BMC Genet 18:12

Lopes M, Bovenhuis H, van Son M, Nordbø Ø, Grindflek E, Knol E et al. (2017) Using markers with large effect in genetic and genomic predictions. J Anim Sci 95:59–71

Luan T, Woolliams J, Lien S, Kent M, Svendsen M, Meuwissen T (2009) The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. Genetics 183:1119–1126

Lukić B, Pong-Wong R, Rowe S, de Koning D, Velander I, Haley C et al. (2015) Efficiency of genomic prediction for boar taint reduction in Danish Landrace pigs. Anim Genet 46:607–616

Lund M, Sahana G, de Koning D, Su G, Carlborg O (2009) Comparison of analyses of the QTLMAS XII common dataset. i: Genomic selection. BMC Proc 3:S1

Meuwissen T (2009) Accuracy of breeding values of unrelated individuals predicted by dense SNP genotyping. Genet Sel Evol 41:35

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

Misztal I, Legarra A (2017) Invited review: efficient computation strategies in genomic selection. Animal 11:731–736

Nejati-Javaremi A, Smith C, Gibson J (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. J Anim Sci 75:1738–1745

Ødegård J, Moen T, Santi N, Korsvoll S, Kjøglum S, Meuwissen T (2014) Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). Front Genet 5:402

Resende MJ, Munoz P, Resende M, Garrick D, Fernando R et al. (2012) Accuracy of genomic selection methods in a standard dataset of Loblolly Pine (Pinus taeda L.). Genetics 190:1503–1510

Schaeffer L (1984) Sire and cow evaluation under multiple trait models. J Dairy Sci 67:1567–1580

Schopen G, Visker M, Koks P, Mullaart E, van Arendonk J, Bovenhuis H (2011) Whole-genome association study for milk protein composition in dairy cattle. J Dairy Sci 94:3148–3158

Sørensen L, Janss L, Madsen P, Mark T, Lund M (2012) Estimation of (co)variances for genomic regions of flexible sizes: application to complex infectious udder diseases in dairy cattle. Genet Sel Evol 44:18

Spindel J, Begum H, Akdemir D, Collard B, Redoña E, Jannink J et al. (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity 116:395–408

Stranden I, Garrick D (2009) Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci 92:2971–2975

Su G, Brøndum R, Ma P, Guldbrandtsen B, Aamand G, Lund M (2012a) Comparison of genomic predictions using medium-density (54,000) and high-density (777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J Dairy Sci 95:4657–4665

Su G, Christensen O, Janss L, Lund M (2014) Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J Dairy Sci 97:6547–6559

Su G, Madsen P, Nielsen U, Mantysaari E, Aaamand G, Christensen O et al. (2012b) Genomic prediction for Nordic Red Cattle using one-step and selection index blending. J Dairy Sci 95:909–917

Taskinen M, Mäntysaari E, Strandén I (2017) Single-step SNP-BLUP with on-the-fly imputed genotypes and residual polygenic effects. Genet Sel Evol 49:36

Thompson R, Meyer K (1986) A review of theoretical aspects in the estimation of breeding values for multi-trait selection. Livest Prod Sci 15:299–313

Tiezzi F, Maltecca C (2015) Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. Genet Sel Evol 47:24

Tsai H, Hamilton A, Tinch A, Guy D, Bron J, Taggart J et al. (2016) Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. Genet Sel Evol 48:47

Wang L, Sørensen P, Janss L, Ostersen T, Edwards D (2013) Genome-wide and local pattern of linkage disequilibrium and persistence of phase for 3 Danish pig breeds. BMC Genet 14:115

Wang Z, Wu Y, Chu H (2018) On equivalence of the LKJ distribution and the restricted Wishart distribution. arXiv e-prints arXiv:1809.04746

Zeng J, Garrick DJ, Dekkers JC, Fernando RL (2016) A nested mixture model for genomic prediction using whole-genome SNP genotypes. Animal Industry Report: AS 662, ASLR3060

Zeng J, Garrick D, Dekkers J, Fernando R (2018) A nested mixture model for genomic prediction using whole-genome SNP genotypes. PLoS ONE 13:e0194683

Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, Zhang Q (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS ONE 5:e12648

Zhou L, Mrode R, Zhang S, Zhang Q, Li B, Liu J (2018) Factors affecting GEBV accuracy with single-step Bayesian models. Heredity 120:100–109