Genomic prediction using multi-trait model with heterogeneous genomic (co)variance

M.T. Anche, G. Su, E. Karaman, L. L. Janss and M. S. Lund

Centre for Quantitative Genetics and Genomics, Aarhus University, DK-8830, Tjele, Denmark

Summary: Single-trait and multi-trait GBLUP models were compared with a recently developed Bayesian model, BayesAS, for their accuracy of genomic prediction. For this purpose, a simulation study were performed considering two correlated traits with heritability of 0.4 and 0.1, various numbers of records for the trait with heritability of 0.1, and three scenarios with regard to varying local and genome-wide (co)variances between the traits. In BayesAS analysis, varying SNP bin sizes defined as a genomic region were used. BayesAS models performed better than GBLUP models across the three scenarios for the high heritability trait in both single-trait and multi-trait analysis. For the trait with lower heritability, the advantage of BayesAS models depended on the number of animals with phenotypic records and the SNP bin size. When increasing the SNP bin size up to 200, multi-trait BayesAS model was able to capture the heterogeneous (co)variance and thus allowed better use of information from correlated trait and increased accuracy of prediction for the trait with low heritability and/or small number of records.
Key words: *genetic correlation, heterogeneous (co)variances, multi-trait genomic prediction, BayesAS model*

Introduction

Most studies on genomic selection have focused on single-trait genomic predictions where only one trait is considered at a time. In most breeding programs, however, individuals are evaluated and selected based on a combination of multiple traits. In addition to that, some of the traits such as feed efficiency and methane emission are difficult and/or expensive to measure and thus are recorded in fewer animals. Currently, with the rapid development of genomic tools, multi-trait genomic prediction has come into play and been applied in a number of studies (Calus and Veerkamp, 2011; Gao *et al.*, 2012; Jia and Jannink, 2012). Compared to single-trait genomic prediction, it was shown that multi-trait genomic prediction can increase the accuracy of genomic prediction considerably for selection candidates without phenotype (Calus and Veerkamp, 2011).

Commonly used multi-trait GBLUP models implicitly assume equal variance and covariance contributed by all the SNPs across the genome. However, this assumption might not be realistic for some traits, where large variance is explained by fewer SNPs with major effects, and the rest contributing a very small proportion. Furthermore, in case of pleiotropic genes, a strong 'local' genomic covariance between SNP effects exists at pleiotropic region while having weak genomic covariance at other regions of the genome. Therefore, failing to include this heterogeneous (co)variance structure across the genome into the genomic models may lead to inaccurate estimates of genomic breeding values, particularly when the genome-wide correlation between the traits is very low, limiting the use of information from correlated traits.

Janss (2014) recently developed a novel Bayesian method, BayesAS, which can model the heterogeneous (co)variance across the genome. This method is expected to yield better estimation of genomic breeding values than those which assume homogenous (co)variance across the genome, as it allows accounting for local specific genetic correlations between the traits. In this study, we aimed to compare the predictive ability of this single-trait (ST-BayesAS) and multi-trait (MT-BayesAS) (Janss, 2014) with single-trait (ST-GBLUP) and multi-trait (MT-GBLUP) genomic best

linear unbiased prediction (GBLUP) models. We also investigated the predictive ability of the BayesAS models with varying SNP bin sizes as a genomic region. The study was carried out using simulated data considering three scenarios with regard to the architecture of genome-wide correlation, SNP bin sizes, and number of phenotypic records.

Materials and Methods

Data

A population of 2,000 Nordic Holsteins cows and 200 bulls genotyped with 50K Illumina BeadChip were used as a base population to simulate 5 non-overlapping generations. Among these SNPs, 200 and 1000 of them were selected to be QTLs. We selected these QTLs based on their MAF (average MAF = 0.15) in such a way that the whole genome was sub-divided into 1million base pairs bins and from these bins we randomly select one bin and choose a marker that has a MAF that is between 0.15±x, where x is a random number from a uniform distribution, *uni*(0, 0.15).

Two traits were simulated with various architectures of genetic correlations across the genome in 3 different scenarios. In scenario 1, 42 QTL were assumed to have the same effect on both traits, and the rest of them affect either of the traits (79 QTL for each trait), leading to a genome-wide genetic correlation of 0.35. Scenario 2 differs from scenario 1 in a way that half of the shared QTL assumed to have same effect on both traits but the other half in opposite direction, leading to a genome-wide genetic correlation of zero. QTL effects for scenarios 1 and 2 were drawn from a gamma distribution with a shape parameter of 0.4, and scale parameter of 1.66 (Hayes and Goddard, 2001; Meuwissen *et al.*, 2001). For scenario 3, 1000 QTL were used and the distribution of the QTL effect were distributed in the same manner as scenario 1 (212 affecting both, 394 QTLs affecting each traits), but the QTL effects were drawn from a normal distribution with 0 mean and variance of 1. In this scenario, genome-wide genetic correlation of 0.35 was assumed.

Breeding values were defined as sum of QTL effects. Residual effects were drawn from the multivariate normal distribution, assuming zero residual correlation. Heritability was 0.4 for trait 1 and 0.1 for trait 2. The cows of generation 0–2 were used as a reference population, while cows of generation 3–5 were used as validation population.

In order to examine the gain from information of a correlated trait in prediction accuracy, varying number of animals with phenotypic records of trait 2 per generation were considered (200, 800, 2,000) while keeping the number of records for trait 1 at 2,000.

Models

The MT-GBLUP model is defined as:

$$\begin{bmatrix}\boldsymbol{y_1}\\ \boldsymbol{y_2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{bmatrix} + \begin{bmatrix}\boldsymbol{Z_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{Z_2}\end{bmatrix}\begin{bmatrix}\boldsymbol{a_1}\\ \boldsymbol{a_2}\end{bmatrix} + \begin{bmatrix}\boldsymbol{e_1}\\ \boldsymbol{e_2}\end{bmatrix} \qquad (1)$$

where $\begin{bmatrix}\boldsymbol{y_1}\\ \boldsymbol{y_2}\end{bmatrix}$ is vector of phenotypic values; $\begin{bmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{bmatrix}$ is vector of overall means; $\begin{bmatrix}\boldsymbol{a_1}\\ \boldsymbol{a_2}\end{bmatrix}$ is vector of random additive genetic effects; $\begin{bmatrix}\boldsymbol{Z_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{Z_2}\end{bmatrix}$ is an incidence matrix linking $\boldsymbol{a}_1$ to $\boldsymbol{y}_1$ and $\boldsymbol{a_2}$ to $\boldsymbol{y_2}$; $\begin{bmatrix}\boldsymbol{e_1}\\ \boldsymbol{e_2}\end{bmatrix}$ is a vector of random residual effect for trait 1 and 2, respectively. It was assumed that

$\begin{bmatrix} \boldsymbol{a_1} \\ \boldsymbol{a_2} \end{bmatrix}$~N(0,$\mathbf{G_0}\otimes\mathbf{G}$), where $\mathbf{G_0}=\begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma^2_{a_2} \end{bmatrix}$, is the variance and covariance matrix of additive genetic effects, and G is the genomic relationship matrix (VanRaden, 2008); and $\begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix}$~N(0, $\mathbf{R}\otimes\mathbf{I}$), where $\mathbf{R}=\begin{bmatrix} \sigma^2_{e_1} & 0 \\ 0 & \sigma^2_{e_2} \end{bmatrix}$, is the variance and covariance matrix for residual effects for trait 1 and 2 and **I** is the identity matrix. The ST-GBLUP model is the simplified version of equation 1 with $\sigma_{a_{12}}= 0$ (Guo *et al.*, 2014).

The MT-BayesAS that uses latent variable to model heterogeneous covariance is defined as:

$$\boldsymbol{y_i} = \mu_i + \sum_{j=1}^{nregion} \boldsymbol{W_j a_{ij}} + \boldsymbol{e_i} \qquad (2)$$

where $\boldsymbol{y_i}$, μ_i, and $\boldsymbol{e_i}$ were the same as model 1, for trait *i*; *nregion* was the total number of genomic regions; $\boldsymbol{W_j}$ is a matrix of SNP genotype covariates (centered) for region *j*, $\boldsymbol{a_{ij}}$ is the vector of SNP effects for trait *i* at region *j*. To handle correlated SNP effect across the traits, the effect of SNP at region *j* was formulated by hierarchical model as follows:

$$a_{ij} = r_i \boldsymbol{s}_{0j} + r_{ij} \boldsymbol{s}_{1j} + a^*_{ij} \qquad (3)$$

where $\boldsymbol{s}_{0j}$ and $\boldsymbol{s}_{1j}$ are latent variables (vectors) and the same for all traits in the model, modelling the overall covariance between traits and regional specific covariance between traits at region *j*, respectively. r_i and r_{ij} are overall and regional specific regression coefficients that determine the size of these covariances. a^*_{ij} are residual SNP effects for trait i at region j. Different sizes of region were investigated, which were, 1-SNP, 50-SNPs, 100-SNPs and 200-SNPs.

The prior distribution for the model unknowns are:

$r_i \sim uni(-\infty, \infty)$, $s_0 \sim N(0, \boldsymbol{I})$, $r_{ij} \sim N(0, \sigma^2_{r_i})$, $\sigma^2_{r_i} \sim uni(0, \infty)$, $s_1 \sim N(0, \boldsymbol{I})$

$a^*_{ij} \sim N(0, \boldsymbol{I}\sigma^2_{a^*_{ij}})$, $\sigma^2_{a^*_{ij}} \sim uni(0, \infty)$

Formulations of variance and covariance for each SNP in each region can be seen in (Li *et al.*, 2017). The ST-BayesAS model is the same as equation 2 but equation 3 reduces to;

$$\boldsymbol{a_{ij}} = r_{ij} * \boldsymbol{s_1} + a^*_{ij} \qquad (4)$$

The prediction accuracies of the above models were calculated as a correlation of estimated and true (simulated) breeding values. The averages of 10 replicates were presented. Due to lack of space, we presented only the results from the following analyses: 1) 2000 records for both traits for all the scenarios, 2) different number of animals having records for trait 2 and different SNP bin sizes for scenario 1.

Results and Discussions

As shown in Table 1, for trait 1, no significant difference in accuracy was observed between scenario 1 and 2 when using both ST and MT models, and when one SNP is considered as a genomic region. This is true for both GBLUP and BayesAS models. For trait 2, on the other hand,

smaller difference in accuracy was observed between scenario 1 and 2, with ST-GBLUP model. When increasing the SNP bin size, however, a small difference in accuracy was observed between scenario 1 and 2 for both traits. Compared to scenario 1 and 2, lower prediction accuracies were observed in scenario 3 when using BayesAS models for both traits.

Because GBLUP models assum equal variances and covariance for all the SNPs, they are unable to capture the existing heterogeneous covariance. As a result, no significant improvement in accuracy was observed when using MT-GBLUP, compared with ST-GBLUP. This can be clearly seen from scenario 2 where there is a zero genome-wide genetic correlation but strong 'local' correlations. As a result, no change in the accuracy was observed moving from ST-GBLUP to MT-GBLUP for both traits.

Within the BayesAS model, the MT model provided substantial increment in accuracy for trait 2 in scenario 1 and 2. The gain in accuracy from multi-trait model in scenario 2 could be attributed to the fact that the model was able to pick up the heterogeneous covariance, since the overall genetic correlation was zero in this scenario. For scenario 3, on the other hand, no (significant) difference in accuracy was observed between ST and MT-BayesAS models for both traits. Moreover, in scenario 3 the BayesAS models provided similar result as the GBLUP models. This could be due to the large number of QTLs with the effect following a normal distribution, thus the BayesAS models work like the GBLUP models.

With regards to using different SNP bin sizes defined as a genomic region, both ST and MT-BayesAS models resulted in substantial increment in prediction accuracy when using bin sizes that are greater than one SNP (Table 1). This was especially true for the trait with lower heritability (trait 2). For trait 1, prediction accuracy increased when using 50 SNPs as a bin size and it plateaued at 100 SNPs and then decreases when 200 SNPs were used as a bin size. This explains that for the trait with higher heritability (trait 1), the effect of a QTL is big enough so that the model can pick the effect efficiently by 50 to 100 SNP around the QTL. For the low heritability trait (trait 2), on the other hand, the QTL effects are smaller compared to trait 1 and thus require large number of SNPs to pick up the effect of a QTL, especially when there are fewer number of individuals with phenotypic records (Table 2). It was observed that one SNP as a region resulted in low genetic correlations between the traits compared to the other cases, and the estimate of genetic correlation increased with increasing the size of bins (result not shown). With one SNP as a genomic region, the model has too many parameters to be estimated and this could also be another reason for the model to be inefficient in genomic prediction.

In Table 2, since only the numbers of individuals with phenotypic records for trait 2 were varied, we presented the result for trait 2 only. When there were 200 individuals with phenotypic records for trait 2, GBLUP models performed as well as and in some cases better than the BayesAS models. When the number of individuals with records increased to 800, both ST and MT BayesAS models provided higher accuracy than the GBLUP models, except for the ST analysis using 1 SNP as a genomic region. This suggests that due to the complication, the BayesAS models may require certain level of information, in order to estimate the parameters accurately and provided higher accuracies than GBLUP models. When 200 SNPs bins were considered as a genomic region, MT-BayesAS model provided a bit higher accuracy for trait 2 than the other bin sizes even when there were fewer records. This suggests that, in order to take advantage of the BayesAS models, especially for a trait with low heritability, determining SNP bin sizes as a genomic region should consider the amount of data information.

Table 1. Accuracy of genomic prediction in different scenarios with 2000 records for both traits

		Scenarios					
		1		2		3	
	Models	Trait 1	Trait 2	Trait 1	Trait 2	Trait 1	Trait 2
Single trait	ST-GBLUP	0.60	0.34	0.60	0.38	0.61	0.40
	ST-BayesAS$_1$*	0.80	0.41	0.80	0.40	0.61	0.40
	ST-BayesAS$_{50}$	0.82	0.60	0.84	0.55	0.64	0.40
	ST-BayesAS$_{100}$	0.82	0.60	0.84	0.57	0.64	0.40
	ST-BayesAS$_{200}$	0.81	0.60	0.82	0.60	0.63	0.40
Multi trait	MT-GBLUP	0.60	0.37	0.60	0.38	0.62	0.42
	MT-BayesAS$_1$	0.80	0.50	0.80	0.50	0.62	0.42
	MT-BayesAS$_{50}$	0.83	0.64	0.84	0.60	0.64	0.42
	MT-BayesAS$_{100}$	0.83	0.64	0.84	0.62	0.64	0.42
	MT-BayesAS$_{200}$	0.81	0.64	0.82	0.62	0.63	0.42

*Sub-indices indicate bin size

Table 2. Accuracy of genomic prediction with different number of records for trait 2 in scenario 1

	# of records	200	800	2000
Single trait	ST-GBLUP	0.15	0.28	0.34
	ST-BayesAS$_1$*	0.13	0.26	0.41
	ST-BayesAS$_{50}$	0.14	0.40	0.60
	ST-BayesAS$_{100}$	0.13	0.36	0.60
	ST-BayesAS$_{200}$	0.14	0.37	0.60
Multi trait	MT-GBLUP	0.21	0.31	0.37
	MT-BayesAS$_1$	0.20	0.36	0.50
	MT-BayesAS$_{50}$	0.21	0.47	0.64
	MT-BayesAS$_{100}$	0.22	0.48	0.64
	MT-BayesAS$_{200}$	0.23	0.50	0.64

*Sub-indices indicate bin size

In this study, we also investigated the decay in accuracy across the three generations of validation population when using GBLUP and BayesAS models (Table 3, from 1 SNP as a genomic region and for the case where all the individuals have phenotypic records for both traits). For scenario 1 and 2, the accuracy decayed more rapidly when using GBLUP models than the BayesAS models. For scenario 3, however, the decay in accuracy was similar for both GBLUP and BayesAS models.

Conclusion

The results of this study showed that accuracy of genomic prediction can be improved using MT model. The BayesAS model allowed accounting for heterogeneous covariance among genomic regions, and thus performed better than GBLUP model. The MT-BayesAS model was able to use information of the other trait as long as regional correlations between traits exist even when genome-wide correlation was zero. The benefit of BayesAS models for a trait with low heritability,

however, depended on the number of records available. In general, the advantage of BayesAS model depended on the distribution of the QTL, the amount of information and SNP bin size used.

Table 3. Decay in accuracy across the validation generation for 1 SNP bin size

		Model	Trait 1			Trait 2		
			Gen 3	Gen 4	Gen 5	Gen 3	Gen 4	Gen 5
Single trait	Scenario 1	GBLUP	0.66	0.61	0.58	0.42	0.37	0.34
		BayesAS	0.78	0.76	0.75	0.46	0.41	0.40
	Scenario 2	GBLUP	0.64	0.58	0.55	0.38	0.34	0.31
		BayesAS	0.80	0.79	0.77	0.44	0.40	0.35
	Scenario 3	GBLUP	0.66	0.61	0.57	0.44	0.39	0.35
		BayesAS	0.65	0.60	0.57	0.42	0.37	0.32
Multi trait	Scenario 1	GBLUP	0.66	0.61	0.58	0.45	0.40	0.37
		BayesAS	0.81	0.78	0.77	0.54	0.53	0.51
	Scenario 2	GBLUP	0.64	0.58	0.55	0.42	0.37	0.34
		BayesAS	0.82	0.81	0.80	0.51	0.48	0.46
	Scenario 3	GBLUP	0.66	0.61	0.57	0.46	0.41	0.38
		BayesAS	0.66	0.61	0.58	0.46	0.41	0.37

References

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

Gao H, Christensen OF, Madsen P, Nielsen US, Zhang Y, Lund MS, *et al.* (2012). Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. *Genet Sel Evol* **44**: 1.

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**: 1–7.

Hayes B, Goddard ME (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genet Sel Evol* **33**: 209–229.

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

L., Janss (2014). Disentangling Pleiotropy along the Genome using Sparse Latent Variable Models. *Proc 10th World Congr Genet Appl to Livest Prod*: 0–3.

Li X, Lund MS, 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**: 26.

Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.

VanRaden PM (2008). Efficient Methods to Compute Genomic Predictions. *J Dairy Sci* **91**: 4414–4423.