A Bayesian Antedependence Model to Account for Linkage Disequilibrium in Whole Genome Selection

Wenzhao Yang*, and *Robert J. Tempelman**

Introduction

Shrinkage based methods, such as BLUP, BayesA, and BayesB (Meuwissen et al., 2001) have been promising in providing a solid statistical foundation for whole genome selection (WGS) on genotypes based on high density SNP marker panels. However, one rather major limitation with all of these methods is the prior specification that all SNP/haplotype effects are in linkage equilibrium (LE) with each other. Inferences on potentially interesting genomic regions can be particularly unstable when extensive linkage disequilibrium (LD) is present due to the redundancy in adjacent markers (Gianola et al., 2009). There has been increasing interest in the use of various geostatistical/longitudinal data analysis tools to accommodate the specification of LD (Gianola et al., 2003; Piepho, 2009). We propose such a method which exploits the strengths of existing Bayesian WGS approaches, based on Markov Chain Monte Carlo (MCMC), while adding very little to their computational load.

Material and methods

Methods Our proposed procedure is based on emerging antedependence models that are being developed for the analysis of longitudinal data (Nunoz-Anton and Zimmerman, 2000) along with recent developments provided in (co)variance modeling (Pourahmadi, 1999). Consider the linear model for WGS modeling on n subjects: $y_i = \mu + \mathbf{z}_i'\mathbf{g} + e_i$; $i = 1,2,\ldots,n$, where μ is an overall mean, $\mathbf{z}_i' = \begin{bmatrix} z_{i1} & z_{i2} & z_{i3} & \ldots & z_{im} \end{bmatrix}$ is a vector of SNP/haplotype-specific covariates on animal i and $\mathbf{g} = \begin{bmatrix} g_1 & g_2 & g_3 & \ldots & g_m \end{bmatrix}'$ is the vector of additive genetic substitution effects for m marker loci. Extensions of this model to accommodate polygenic effects, due to insufficient marker coverage, are possible and generally necessary following Habier et al. (2007); however, we do not pursue this further for pedagogical reasons.

Suppose that the subscripts of the elements of $\mathbf{g}$ specify the relative order of the SNPs such that the following antedependence structure is considered: $g_1 = \delta_1$, $g_2 = t_{21}g_1 + \delta_2$, $g_3 = t_{32}\,g_2 + \delta_3$, ... , $g_m = t_{m,m\text{-}1}\,g_{m\text{-}1} + \delta_m$. The vector $\mathbf{t} = \begin{bmatrix} t_{21} & t_{32} & t_{43} & \ldots & t_{m,m-1} \end{bmatrix}'$ thereby specifies correlations between effects at SNPs because of LD using a heterogeneous first order antedependence structure. For markers specifying demarcations between different linkage groups, the corresponding elements of $\mathbf{t}$ would be set to 0. Furthermore, assume that

* Michigan State University, East Lansing, MI, USA, 48824-1225

elements of $\boldsymbol{\delta} = \begin{bmatrix} \delta_1 & \delta_2 & \delta_3 & \dots & \delta_m \end{bmatrix}'$ are specified to be normally, and independently distributed with null mean and SNP specific variances: $\boldsymbol{\Delta} = \begin{bmatrix} \sigma^2_{\delta 1} & \sigma^2_{\delta 2} & \sigma^2_{\delta 3} & \dots & \sigma^2_{\delta m} \end{bmatrix}'$.

Writing **T** as a *m* x *m* matrix, being all 0's except for the elements with corresponding subscript addresses specified above in **t**, we can re-express the above relationships as $\mathbf{g} = \mathbf{Tg} + \boldsymbol{\delta}$ such that $\mathbf{G} = \text{var}(\mathbf{g}) = \mathbf{T\Delta T}'$. Note that elements of **t** can be unconstrained such that **G** is guaranteed to be positive semi-definite. Furthermore, although **G** is dense, $\mathbf{G}^{-1} = (\mathbf{I}-\mathbf{T})'\boldsymbol{\Delta}^{-1}(\mathbf{I}-\mathbf{T})$ is a sparse block tridiagonal matrix with one block per linkage group. This is significant given that $\mathbf{G}^{-1}$, rather than **G**, is required for specifying full conditional densities (FCD) for elements of $\boldsymbol{\delta}$ (and hence **g**) based on MCMC implementations for WGS models as in Meuwissen et al. (2001).

We further specify that the elements of **t** are independent random draws from $N(\mu_t,\sigma^2_t)$ where μ_t and σ^2_t are, in turn, are unknown and, hence, assigned vaguely informative priors. It can be subsequently shown that the FCD of the elements of **t** are also normal. We further specify that the elements of $\boldsymbol{\Delta}$ are independently distributed as random draws from a scaled inverted chi-square distribution, $\chi^{-2}(\nu,S)$, similar in spirit to the BayesA implementation of Meuwissen et al. (2001). With everything else being identical except for the antedependence structure on **g**, we call our model *Ante-BayesA*. We could similarly specify a mixture prior (e.g., $\sigma^2_{\delta j} = 0$ with probability π and $\sigma^2_{\delta j} \sim \chi^{-2}(\nu,S)$ with probability 1- π for j=1,2,…,m) to develop a competitor method with BayesB (Meuwissen et al., 2001).

Simulation study. A small-scale simulation study was created to study the relative merit of using Ante-BayesA to accommodate LD. The genome consisted of a single 100cM chromosome with 10001 SNP markers and 10000 potential QTLs, one QTL per each SNP marker interval. Mutation rates for markers and QTLs were specified to be $2.5\text{x}10^{-3}$ and $2.5\text{x}10^{-4}$, respectively, for each of 100 individuals per each of 6000 generations; furthermore, mutation was specified to be recurrent between only two possible biallelic states. Upon mutation, QTL effects were generated from a Gamma(0.4,1.66) distribution with the sign of the effect randomly assigned to be either positive or negative with equal probabilities as in Meuwissen et al. (2001) . Phenotypes from the last generation were generated by scaling the sum of the QTL effects to have a variance of 1 and adding to that a $N(0,1)$ residual such that the heritability of the response was 0.50. Different levels of LD (r^2 = 0.19-0.24) were generated by randomly choosing 1000 SNP markers from those exceeding different thresholds on minor allele frequencies (MAF), such that choosing higher MAF thresholds led to higher LD levels. Both Meuwissen's Bayes A and our proposed modification, Ante-BayesA, were compared for accuracy of selection, defined as the correlation between true breeding value (TBV = $\mathbf{z}_i'\mathbf{g}$) and estimated breeding value (EBV = $\mathbf{z}_i'\hat{\mathbf{g}}$) based on the posterior means ($\hat{\mathbf{g}}$) of **g**. The simulation study was replicated 10 times.

It is important to realize with this simulation study that it is not possible to know the true values of **t;** their elements are merely proxies for modeling LD dependencies between SNPs and QTLs.

Results and discussion

MCMC mixing was generally much better using the Ante-BayesA method compared to the BayesA method, particularly with higher LD levels. This result is likely because the elements of $\boldsymbol{\delta}$ corresponding to adjacent SNPs have far less posterior correlation with each other compared to the corresponding elements of **g.** That is, the Ante-BayesA procedure explicitly avoids marker redundancy by conditioning SNP genetic effects on effects of the neighbouring SNP. The comparisons between BayesA and Ante-BayesA for accuracy (correlation between TBV and EBV) across the 10 replicates is summarized in Figure 1. At lower levels of LD ($r^2 \leq 0.20$), there was no statistically significant evidence of a difference in accuracy between the two methods, whereas Ante-BayesA showed substantially improved accuracies for higher levels of LD. These results are expected given that lower levels of LD will increasingly agree with the LE-based assumptions of Meuwissen's BayesA whereas higher levels of LD will, at least partly, be accommodated by the first-order antedependence assumptions of Ante-BayesA. We also found that the accuracy of selection for the Ante-BayesA procedure was invariant to either of the two directions that the antedependence structure was specified along the chromosome.

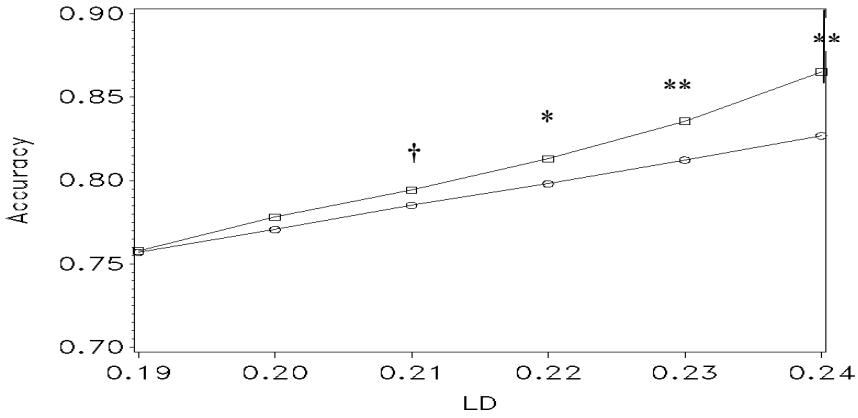

Figure 1: Accuracy of breeding value prediction for Bayes A(○○○) and Ante Bayes A (□□□) . Differences in accuracy significant at P<.05 (†), P<.001 (*), and P<.0001 (**).

For each LD scenario and replicate, we also investigated the proportion of SNP intervals where the corresponding posterior means of elements of **t** divided by their respective posterior standard deviations exceeded 2 (i.e., leading to an approximate two-tailed Bayesian *p*-value of 0.05); the greater this proportion, the greater the amount of LD is being captured. For example, if this proportion is 0.50 for a particular replicate, that implies that 50% of the elements of **t** are statistically significant. A boxplot of this statistic for each SNP interval across the 10 replicates for each of the LD levels is specified in Figure 2. Quite clearly, the importance of the elements of **t** is greater with higher levels of LD. Absolute elements of **t** were, in particular, large in close proximity to QTLs of large effect.

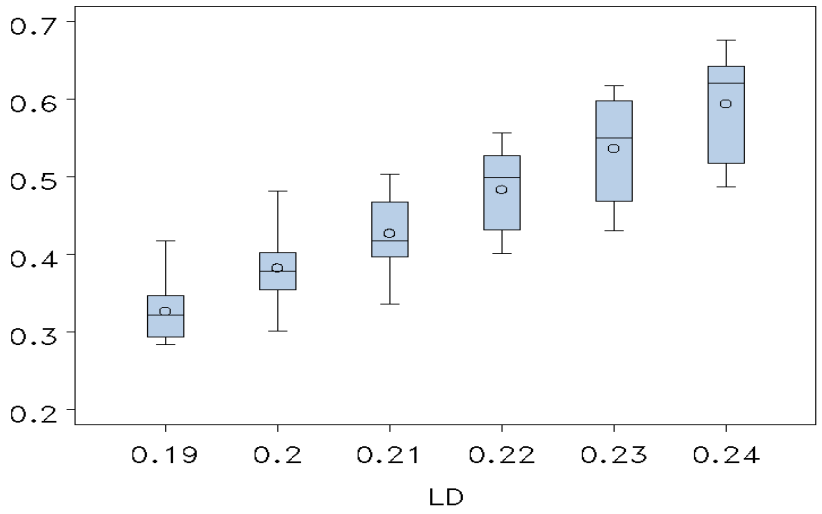

Figure 2: Boxplot of proportions of times that the absolute values of the posterior means of elements of t divided by their respective posterior standard deviations exceeded 2 across replicates for each LD level.

Conclusion

We believe that we have developed a biologically reasonable and computationally tractable method to accommodate LD based on existing hierarchical Bayesian WGS modeling without haplotyping. This antedependence based model should lead to measurably greater gains in accuracy of WGS as greater levels of LD are attained between markers with newly developed SNP marker panels.

References

Gianola, D., Pérez-Enciso, M., and Toro, M.A. (2003). *Genetics,* 163:347-365.

Gianola, D., de los Campos, G., Hill, W.G. *et al.* (2009). *Genetics*, 183:347-363.

Habier, D., Fernando, R.L., Dekkers, J.C.M. (2007) *Genetics*, 177:2389-2397.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). *Genetics,* 157:1819–1829.

Nunoz-Anton, V., Zimmerman, D.L. (2000) *Biometrics,* 56:699-705.

Piepho, H.P. (2009). *Crop Sci*., 49:1165-1176.

Pourahmadi, M. (1999). *Biometrika* 86:677-690.