J. Dairy Sci. 97:1–13
http://dx.doi.org/10.3168/jds.2014-8210

Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances

G. Su,[1] O. F. Christensen, L. Janss, and M. S. Lund
Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark

ABSTRACT

Various models have been used for genomic prediction. Bayesian variable selection models often predict more accurate genomic breeding values than genomic BLUP (GBLUP), but GBLUP is generally preferred for routine genomic evaluations because of low computational demand. The objective of this study was to achieve the benefits of both models using results from Bayesian models and genome-wide association studies as weights on single nucleotide polymorphism (SNP) markers when constructing the genomic matrix (G-matrix) for genomic prediction. The data comprised 5,221 progeny-tested bulls from the Nordic Holstein population. The animals were genotyped using the Illumina Bovine SNP50 BeadChip (Illumina Inc., San Diego, CA). Weighting factors in this investigation were the posterior SNP variance, the square of the posterior SNP effect, and the corresponding minus base-10 logarithm of the marker association P-value $[-\log_{10}(P)]$ of a t-test obtained from the analysis using a Bayesian mixture model with 4 normal distributions, the square of the estimated SNP effect, and the corresponding $-\log_{10}(P)$ of a t-test obtained from the analysis using a classical genome-wide association study model (linear regression model). The weights were derived from the analysis based on data sets that were 0, 1, 3, or 5 yr before performing genomic prediction. In building a G-matrix, the weights were assigned either to each marker (single-marker weighting) or to each group of approximately 5 to 150 markers (group-marker weighting). The analysis was carried out for milk yield, fat yield, protein yield, fertility, and mastitis. Deregressed proofs (DRP) were used as response variables to predict genomic estimated breeding values (GEBV). Averaging over the 5 traits, the Bayesian model led to 2.0% higher reliability of GEBV than the GBLUP model with an original unweighted G-matrix. The superiority of using a GBLUP with weighted G-matrix over GBLUP with an original unweighted G-matrix was the largest when using a weighting factor of posterior variance, resulting in 1.7 percentage points higher reliability. The second best weighting factors were $-\log_{10}$ (P-value) of a t-test corresponding to the square of the posterior SNP effect from the Bayesian model and $-\log_{10}$ (P-value) of a t-test corresponding to the square of the estimated SNP effect from the linear regression model, followed by the square of estimated SNP effect and the square of the posterior SNP effect. In addition, group-marker weighting performed better than single-marker weighting in terms of reducing bias of GEBV, and also slightly increased prediction reliability. The differences between weighting factors and scenarios were larger in prediction bias than in prediction accuracy. Finally, weights derived from a data set having a lag up to 3 yr did not reduce reliability of GEBV. The results indicate that posterior SNP variance estimated from a Bayesian mixture model is a good alternative weighting factor, and common weights on group markers with a size of 30 markers is a good strategy when using markers of the 50,000-marker (50K) chip. In a population with gradually increasing reference data, the weights can be updated once every 3 yr.

Key words: genomic relationship matrix, genomic selection, model, reliability

INTRODUCTION

Several statistical models have been proposed for genomic predictions using genome-wide SNP markers. One of the most popularly used models is genomic BLUP (**GBLUP**), which is a linear mixed model incorporating a marker-based genomic relationship matrix (**G-matrix**), because it is in the same form as a simple traditional BLUP model and has a low computational requirement. The G-matrix is built using the information of genome-wide dense markers (VanRaden, 2008; Hayes et al., 2009b). Compared with traditional pedigree-based models, it has the advantage of being able to capture linkage disequilibrium (**LD**) between markers and causal genes, Mendelian segregation, and genetic links through unknown common ancestors that are not available in the known pedigree. Therefore, the G-matrix is superior to the pedigree-based relationship

Received April 6, 2014.
Accepted July 7, 2014.
[1]Corresponding author: guosheng.su@agrsci.dk

matrix for genetic evaluation, and can be implemented in methods and models that conventionally incorporate a pedigree-based relationship matrix. In GBLUP models, the covariance matrix of additive genetic effect is generally defined to be proportional to the G-matrix under the assumption that given the same allele frequency, all the markers have the same contribution to the genetic variation of a trait. This is equivalent to assuming that the effects of all markers follow the same normal distribution (VanRaden, 2008; Strandén and Garrick, 2009). Obviously, the assumption is not desirable if the trait is affected by major genes.

Unlike GBLUP, Bayesian variable selection models allow variances of SNP effects differing among loci. This is usually realized by assuming a thick-tailed distribution of SNP effects or a mixture of 2 or more distributions (Meuwissen et al., 2001; Su et al., 2010; Habier et al., 2011; Erbe et al., 2012; Gao et al., 2013). Clearly, Bayesian variable selection models capture the feature of SNP effects better than GBLUP. Many simulation studies have shown that Bayesian models perform better than the GBLUP model (Meuwissen et al., 2001; Lund et al., 2009; Guo et al., 2010). Based on real cattle data, previous studies showed that Bayesian models led to similar or higher prediction accuracies than GBLUP (Hayes et al., 2009a; Habier et al., 2010; Su et al., 2012a). The benefit from Bayesian models is larger for traits controlled by large QTL (Cole et al., 2009; Legarra et al., 2011) and for animals that have weak relationship with individuals in the reference population (Habier et al., 2010; Gao et al., 2013). However, typical Bayesian variable selection models using the Markov chain Monte Carlo (**MCMC**) algorithm have the disadvantage of a long computing time.

Because in a GBLUP model the covariance matrix is proportional to the G-matrix, a simple way to overcome the disadvantage of the inappropriate assumptions of an original GBLUP model is to build a G-matrix in which markers are weighted appropriately. Zhang et al. (2010) proposed a method to build a trait-specific G-matrix for genomic prediction using GBLUP and analyzed simulation data. They reported that the accuracy of genomic EBV (**GEBV**) increased when using a GBLUP model with a G-matrix weighted by squared marker effect estimated from a random regression BLUP model, compared with the original GBLUP. The accuracies further increased when using a G-matrix weighted with the posterior variance of the marker effect estimated from a BayesB model. Based on data from French Holstein and Montbéliarde cattle, Legarra et al. (2011) predicted genomic breeding values using a GBLUP with a G-matrix weighted by the posterior variance of the marker effect derived from the analysis using a Bayesian least absolute shrinkage and selection operator (LASSO) model, and obtained prediction accuracies close to those using the Bayesian LASSO directly. A study by de los Campos et al. (2013) used the minus base-10 logarithm of the marker association *P*-value [$-\mathbf{log_{10}}(\boldsymbol{P})$] from a genome-wide association study (**GWAS**) as weight on the makers to build a G-matrix, and reported that the weighted G-matrix improved prediction accuracy, based on human type-2 diabetes case-control data sets. In contrast, Zhou et al. (2014) reported that a G-matrix weighted with the square of estimated marker effect led to lower accuracy of genomic prediction than an original G-matrix, based on the Nordic dairy cattle data.

It can be hypothesized that an appropriately weighted G-matrix can improve the prediction reliability of a GBLUP model, and a GBLUP model with a G-matrix weighted using the posterior variance of the marker effect from a Bayesian variable selection model can achieve the similar prediction reliability as the Bayesian variable selection model. The objective of this study was to test these hypotheses by assessing alternative weighting factors to construct weighted G-matrices for genomic prediction. In addition, this study investigated the ways to weight markers, and the time intervals when weights need to be updated. The analysis was based on data from the Nordic Holstein population.

MATERIALS AND METHODS

Data

The data in this analysis comprised 5,643 progeny-tested bulls from the Nordic Holstein population. The animals were genotyped with the Illumina Bovine SNP50 BeadChip (Illumina Inc., San Diego, CA; Matukumalli et al., 2009). The marker data were edited by deleting markers with minor allele frequency lower than 0.01, average GenCall score lower than 0.60, or unknown location in the UMD 3.1 assembly [University of Maryland, College Park]. After editing, 44,919 markers remained in the analysis. Sporadic missing genotypes were imputed using the BEAGLE software package (Browning and Browning, 2009). The phenotypic data for genomic prediction were deregressed proofs (**DRP**), which were derived from the Nordic genetic evaluations in January 2013. The traits in the analysis were milk yield, fat yield, protein yield, fertility, and mastitis.

The data were divided into a reference data set and a validation data set by birth date (January 1, 2005). This resulted in about the 20% youngest bulls being validation bulls. Deregressed proofs with reliability less than 10% were excluded from the reference data and less than 20% were removed from the test data. The number of animals with phenotypic information dif-