J. Dairy Sci. 93:1175–1183
doi:10.3168/jds.2009-2192

Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population

G. Su,[1] B. Guldbrandtsen, V. R. Gregersen, and M. S. Lund
Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, Aarhus University, DK-8830, Tjele, Denmark

ABSTRACT

This study investigated the reliability of genomic estimated breeding values (GEBV) in the Danish Holstein population. The data in the analysis included 3,330 bulls with both published conventional EBV and single nucleotide polymorphism (SNP) markers. After data editing, 38,134 SNP markers were available. In the analysis, all SNP were fitted simultaneously as random effects in a Bayesian variable selection model, which allows heterogeneous variances for different SNP markers. The response variables were the official EBV. Direct GEBV were calculated as the sum of individual SNP effects. Initial analyses of 4 index traits were carried out to compare models with different intensities of shrinkage for SNP effects; that is, mixture prior distributions of scaling factors (standard deviation of SNP effects) assuming 5, 10, 20, or 50% of SNP having large effects and the others having very small or no effects, and a single prior distribution common for all SNP. It was found that, in general, the model with a common prior distribution of scaling factors had better predictive ability than any mixture prior models. Therefore, a common prior model was used to estimate SNP effects and breeding values for all 18 index traits. Reliability of GEBV was assessed by squared correlation between GEBV and conventional EBV ($r^2_{GEBV, EBV}$), and expected reliability was obtained from prediction error variance using a 5-fold cross validation. Squared correlations between GEBV and published EBV (without any adjustment) ranged from 0.252 to 0.700, with an average of 0.418. Expected reliabilities ranged from 0.494 to 0.733, with an average of 0.546. Averaged over 18 traits, $r^2_{GEBV, EBV}$ was 0.13 higher and expected reliability was 0.26 higher than reliability of conventional parent average. The results indicate that genomic selection can greatly improve the accuracy of preselection for young bulls compared with traditional selection based on parent average information.

Key words: cross validation, genomic estimated breeding value, genomic selection, reliability

INTRODUCTION

The application of molecular genetic information has become an important issue in animal breeding. In cattle, an assay for simultaneous genotyping of more than 50,000 SNP markers is commercially available. This opens an opportunity for effective selection using dense markers through the whole genome (i.e., genomic selection). Genomic selection is based on breeding values that are directly estimated from genome-wide dense marker panels. Therefore, genetic evaluation can be performed as soon as DNA is obtained, which allows accurate selection in both genders early in life. Genomic selection is expected to lead to considerably higher genetic gains than conventional quantitative genetic selection (Meuwissen et al., 2001; Schaeffer, 2006). It is expected that by using genomic selection in dairy cattle breeding, the genetic progress would be doubled whereas the cost for proving bulls would be reduced by 92% (Schaeffer, 2006).

Several statistical models and algorithms have been proposed to predict breeding values based on dense markers (Meuwissen et al., 2001; Xu, 2003; Meuwissen and Goddard, 2004; Gianola et al., 2006). Among the proposed methods, BLUP, BayesA, and BayesB have been widely used to analyze simulated data and real data. A linear BLUP approach (Meuwissen et al., 2001; VanRaden, 2008) assumes that effects of all SNP are normally distributed with same variance. BayesA and BayesB (Meuwissen et al., 2001) allow each marker to have its own variance of allele effects, and each variance is a sample of a scaled inverse chi-squared distribution. BayesB also models most SNP having zero effect, but a few having moderate to large effects. To simplify the computing algorithm in BayesA and BayesB (especially the Metropolis-Hastings step in BayesB), alternative Bayesian approaches similar to BayesA and BayesB have been proposed for genomic prediction (Meuwissen and Goddard, 2004; Villumsen et al., 2009). These approaches model SNP effects as a product of a scaled effect and a scaling factor (which can be understood

Received March 6, 2009.
Accepted December 3, 2009.
[1]Corresponding author: guosheng.su@agrsci.dk

as standard deviation of allele effects in a marker). It assumes that the prior distribution of scaling factors is either a normal distribution or a mixture of 2 normal distributions. Some simulation studies showed that the prediction ability of BayesA and BayesB was greater than BLUP approach, based on the simulated scenarios assuming that few QTL had a large effect and most QTL had a small effect (Meuwissen et al., 2001; Lund et al., 2009). Based on real data from dairy cattle, the Bayesian approaches or the analogous approaches [e.g., nonlinear BLUP in VanRaden (2008)] gave higher reliabilities than linear BLUP approach for the traits having known major genes, but the differences between these approaches were very small for the traits without major genes (Cole et al., 2009; Hayes et al., 2009; VanRaden et al., 2009).

So far, most reports on genomic selection in the literature were based on simulated data. Recently, many results based on the data from real livestock populations have been published (e.g., Harris et al., 2008; González-Recio et al., 2009; Hayes et al., 2009; VanRaden et al., 2009). However, to apply this new technology in practical breeding programs, it is necessary to evaluate the accuracy of genomic prediction in the target population. Therefore, the objectives of this study were to assess the predictive ability of models with different prior densities of marker effects and to investigate the reliability of genomic estimated breeding values for 18 traits based on the data from the Danish Holstein population.

MATERIALS AND METHODS

Data

Holstein bulls from 258 half-sib families (1–41 bulls each), born between 1986 and 2004, were genotyped using Illumina Bovine SNP50 BeadChip (Illumina, San Diego, CA). The marker data were edited using the following procedures: 1) deleted the locus with minor allele frequency less than 5%; 2) deleted the locus with average GenCall score less than 0.65; 3) deleted the individual with call rate score less than 0.85; and 4) for a marker with GenCall score less than 0.6 in an individual, set the marker as unknown in this individual. After the editing, there were 3,330 bulls and 38,134 SNP markers available. In the analysis of SNP effects and genomic prediction, any missing SNP at a particular marker in some animals was treated as an extra allele. This corresponded to replacing the effect of missing SNP at a marker with population mean of this marker.

Published conventional EBV were used as response variables to estimate SNP effects. The EBV and their reliability for the genotyped bulls were obtained from official evaluations in April 2009. In total, 18 index traits were analyzed in this study. Except for fat percentage and protein percentage, the traits are the subtraits in the new Nordic Total Merit index. Detailed descriptions of these index traits and their EBV are given in Danish Cattle Federation (2006).

Statistical Model

In this study, all individual SNP markers were used as predictors and conventional EBV were used as response variables weighted by a function of reliability of EBV (see detail later). A Bayesian method, which captures the features of BayesA and BayesB but simplifies the computing algorithm, was used to estimate marker effects for genomic prediction. The method applies the methodology of variable selection presented by George and McCulloch (1993). A detailed description of the method was presented by Villumsen et al. (2009) and Meuwissen and Goddard (2004). The following model was used to fit EBV data:

$$\mathbf{y} = 1\mu + \sum_{\mathrm{i=1}}^{\mathrm{m}} \mathbf{X}_i \mathbf{q}_i \nu_i + \mathbf{e},$$

where $\mathbf{y}$ is the vector of published conventional EBV, μ is the intercept, m is the number of SNP markers, $\mathbf{X}_i$ is the design matrix of allele types in marker i, $\mathbf{q}_i$ is the vector of scaled SNP effects (scaled by SD) of marker i with $\mathbf{q}_i \sim N(\mathbf{0}, \mathbf{I})$, v_i ($v_i > 0$) is a scaling factor (SD) for SNP effects of marker i, and $\mathbf{e}$ is the vector of residual with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where I is a identity matrix. The effects of SNP alleles of marker i are the products of v_i and $\mathbf{q}_i$.

Scaling factors v_i were assumed to have either a common prior distribution or a mixture prior distribution. A common prior distribution across the variances of chromosome segment effects, which leads to a slight or moderate differentiation between small and large effects of markers, was assumed to be a positive half-normal distribution (TN),

$$v_i \sim TN(0, \sigma_v^2),$$

where $v_i > 0$. Mixture prior distributions, which lead to strong differentiation between small and large effects of markers, assume that a proportion (π_0, typically large) of markers have very small effects, and another proportion ($\pi_1 = 1 - \pi_0$, typically small) of markers have moderate or large effects. This was achieved by assuming that the prior distribution of v_i was sampled from