J. Dairy Sci. 98:2775–2784
http://dx.doi.org/10.3168/jds.2014-8975

Single-step genomic evaluation using multitrait random regression model and test-day data

M. Koivula,*[1] I. Strandén,* J. Pösö,† G. P. Aamand,‡ and E. A. Mäntysaari*
*Natural Resources Institute Finland (Luke), Green Technology, 31600 Jokioinen, Finland
†Faba Co, 01301 Vantaa, Finland
‡NAV Nordic Cattle Genetic Evaluation, Agro Food Park 15, 8200 Aarhus N, Denmark

ABSTRACT

The objectives of this study were to evaluate the feasibility of use of the test-day (TD) single-step genomic BLUP (ssGBLUP) using phenotypic records of Nordic Red Dairy cows. The critical point in ssGBLUP is how genomically derived relationships (**G**) are integrated with population-based pedigree relationships (**A**) into a combined relationship matrix (**H**). Therefore, we also tested how different weights for genomic and pedigree relationships affect ssGBLUP, validation reliability, and validation regression coefficients. Deregressed proofs for 305-d milk, protein, and fat yields were used for a posteriori validation. The results showed that the use of phenotypic TD records in ssGBLUP is feasible. Moreover, the TD ssGBLUP model gave considerably higher validation reliabilities and validation regression coefficients than the TD model without genomic information. No significant differences were found in validation reliability between the different TD ssGBLUP models according to bootstrap confidence intervals. However, the degree of inflation in genomic enhanced breeding values is affected by the method used in construction of the **H** matrix. The results showed that ssGBLUP provides a good alternative to the currently used multi-step approach but there is a great need to find the best option to combine pedigree and genomic information in the genomic matrix.
Key words: genomic evaluation, single step, test-day model, Nordic Red Dairy cow, single-step genomic BLUP (ssGBLUP)

INTRODUCTION

Most genomic evaluations are based on multi-step approach that requires (1) calculation of traditional EBV without genomic information; (2) extraction of pseudo-observations, typically either daughter yield deviations (**DYD**) or deregressed EBV (deregressed proofs; **DRP**); and (3) genomic model for prediction of direct genomic values (**DGV**; VanRaden, 2008; Hayes et al., 2009; VanRaden et al., 2009). Genomic evaluations can be further improved by combining the DGV and information from traditional EBV (e.g., VanRaden, 2008) to yield genomic enhanced breeding values (**GEBV**).

The multi-step approach to calculate GEBV has an inherent problem. First, the parent averages (**PA**) of progeny of genomically selected animals do not automatically include genomic information. Second, when animals are selected by their GEBV, the future estimation of unbiased EBV becomes difficult because genomic information is not taken into account in the traditionally calculated EBV. Moreover, genomic selection using the multi-step approach is complex and includes several approximations, all of which reduce accuracy and can inflate the resultant GEBV. None of these issues applies to the single-step approach.

Single-step evaluation (single-step genomic BLUP; **ssGBLUP**) is a unified approach to calculate GEBV. The ssGBLUP combines phenotypic records, pedigree information, and genomic information optimally in calculation of GEBV (Misztal, et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010). The approach integrates the pedigree relationship matrix **A** and genomic relationship matrix **G** into a single **H** matrix, which replaces the traditional relationship matrix **A** in the mixed-model equations (Legarra et al., 2009; Misztal et al., 2009; Christensen and Lund, 2010). To date, the single-step approach has been rated computationally demanding with large data sets and multi-trait analysis (Su et al., 2012). However, ssGBLUP has been successfully applied, for example, for final scores of over 6 million Holsteins with greater accuracy than that of a multi-step procedure (Aguilar et al., 2010), and in a multi-trait national genomic evaluation for type traits in US Holsteins (Tsuruta et al., 2011). Performance of ssGBLUP has also been evaluated in other species. Chen et al. (2011) used ssGBLUP to analyze 3 traits in 2 separately selected lines of chickens, and Forni et

Received October 14, 2014.
Accepted December 16, 2014.
[1]Corresponding author: minna.koivula@luke.fi

al. (2011) used ssGBLUP to analyze litter size in pigs. Thus, despite its high computational requirements, the single-step method is suitable for large multi-trait analyses. The critical issue with ssGBLUP is compatibility between the marker-based relationship matrix and the pedigree-based relationship matrix for genotyped animals. Applications of the first unified approaches for merging information from animals with or without genotypes by combining the **A** matrix with the **G** matrix resulted in biased GEBV (e.g., Meuwissen et al., 2011). Since then, it has been demonstrated that accuracy of prediction can be improved and bias reduced by adjusting the **G** matrix toward their expected values in the **A** matrix to decrease the scaling problem (e.g., Vitezica et al., 2011; Christensen et al., 2012).

A random regression test-day (**TD**) model is currently used for the official Nordic genetic evaluation of production (Lidauer et al., 2015) in Nordic Red Dairy Cattle (**RDC**). As more selection decisions are made using genomic information, it is becoming essential that all genomic information is included in national evaluations. The objectives of this study were to evaluate the feasibility of the large random regression TD ssGBLUP, and to estimate the accuracy of GEBV when using this model. We also tested how different combinations of the **A** and **G** matrices affect the bias and accuracy of GEBV in the TD ssGBLUP.

MATERIALS AND METHODS

All analyses used the data used in the official Nordic RDC milk production evaluations. The multiple-trait milk production evaluation includes TD records for milk, fat, and protein production. Production records from the first 3 lactations are in the same multiple-trait model. Each trait has random regression function for random genetic and permanent environmental effects. For more information, see Lidauer et al. (2015).

The routine full evaluation data from May 2014 for the RDC were obtained from the Nordic Cattle Genetic Evaluation (**NAV**; Aarhus, Denmark). For production traits, the TD data included 3.8 million cows with a total of 85 million records and 5.1 million animals in the Nordic RDC pedigree. To be able to validate the model, a reduced data set was extracted from the full data set, as follows: the last 4 yr of observations were removed and the reduced data included 2.7 million cows with 72 million records. The reduced data set was used to solve GEBV and EBV for all animals in the pedigree, and the full data set was used to solve current EBV for testing purpose. The initial EBV from the reduced data set were denoted $\mathbf{EBV_r}$. For the females without observations and bulls without daughters in reduced data, EBV_r are hereafter referred to as parent average (**PA**). Comparing initial predictions from the reduced data set with those from the full data set allowed estimation of validation accuracy (Mäntysaari et al., 2010). The total number of equations in the reduced run was 217,370,251, and in the full run 238,041,030.

The unified relationship matrix **H** in single-step evaluations defines the relationships among genotyped and nongenotyped animals. Although **H** can be expensive to compute, its inverse has a simple structure (Aguilar et al., 2010; Christensen and Lund, 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{A}_{22}$ is the sub-matrix of the pedigree-based numerator relationship matrix **A** for the genotyped animals, and **G** is the relationship matrix constructed using genomic information. The **G** matrix had 15,148 genotyped RDC animals, of which 5,534 were bulls and 9,529 cows. The **G** matrix also included genotypes of animals without offspring or records. Genotypes were obtained from the Illumina Bovine SNP50 Bead Chip (Illumina, San Diego, CA). After application of exclusion criteria, 46,914 SNP markers on the 29 bovine autosomes were available for further analysis. The genotype file was the same as was used in official genomic evaluation of Nordic Cattle Genetic Evaluation in June 2014. Genotypes were used to form the raw **G** matrix with method 1 in VanRaden (2008). Before the matrices **G** and $\mathbf{A_{22}}$ were combined, the raw **G** matrix was scaled by scalar $t = \frac{tr(\mathbf{A}_{22})}{tr(\mathbf{G})}$, where tr is the trace of matrix. Thus, **G** has, on average, the same diagonals as the $\mathbf{A}_{22}$ matrix.

When the mixed-model equation for single-step is considered, the difference from the normal animal model is the matrix block $\mathbf{H}^{22} = \mathbf{A}^{22} + \mathbf{G}^{-1}–\mathbf{A_{22}}^{-1}$ between genotyped animals. To improve the properties of the ssGBLUP, different weights in building the $\mathbf{H}^{22}$ matrix were tested. Aguilar et al. (2010) and Christensen and Lund (2010) noted that if not all genetic variance is accounted for by the SNP effects, the residual polygenic effect can be included in the model by changing the genomic matrix **G** and using $\mathbf{H}^{22} = \mathbf{A}^{22} + \mathbf{G}_{w}^{-1} - \mathbf{A}_{22}^{-1}$, where $\mathbf{G}_{w} = (1 - w)\mathbf{G} + w\,\mathbf{A}_{22}$, and the constant w represents the proportion of polygenic variance not described by markers. So, the smaller w, the more genetic variance that is attributed to genomic markers. We used 3 different proportions w ($w = 0.10$, $w = 0.15$, or $w = 0.20$) in $\mathbf{G}_{w}$. In Christensen et al. (2012), the optimal w was found to be 0.20, al-