311. An alternative method to consider a reference population in Single-Step SNP BLUP model without separating genomic terms

T. Osawa[1*], T. Baba[2], Y. Masuda[3], T. Kawahara[2] and Y. Goto[2]

[1]National Livestock Breeding Center, Nishigo-mura, Fukushima, 961-8511, Japan; [2]Holstein Cattle Association of Japan, Hokkaido Branch, Sapporo, Hokkaido, 001-0015, Japan; [3]Rakuno Gakuen University, Ebetsu, Hokkaido, 069-8501, Japan; t0ohsawa@nlbc.go.jp

Abstract

This study proposed an alternative method to consider a reference population in the mixed model equations (MME) of single-step SNP BLUP (ssSNP-BLUP) model without separating genomic terms to estimate SNP effects. We adopted the preconditioned biconjugate gradient stabilized method to solve MME because this model was a nonsymmetric linear system. Overall conformation score in the first lactation for Japanese Holsteins was used to compare the original and modified ssSNP-BLUP models. Genotyped animals included in the dataset were 37,197 cows, 5,352 sires, 3,973 young bulls and 88,058 heifers. Three reference populations were defined as sires, cows and both. Our method can consider the reference population in the ssSNP-BLUP without separating genomic terms. The contribution of young animals to genomic predictions was small, and predictions varied when only sires or cows were considered in a reference population. On the other hand, considering the reference population may reduce the overestimation of the predictions.

Introduction

Currently, genetic evaluations of dairy cattle in many countries include large-scale SNP marker information, known as genomic evaluation. Most countries have applied a multi-step model for the genomic evaluation of dairy cattle (VanRaden., 2008), which utilizes deregressed estimated breeding values (EBV) obtained via prior conventional genetic evaluation based on pedigree information. The multi-step model is feasible to implement in conventional genetic evaluation systems. A major advantage of the multi-step method is its low computational cost. It uses only selected genotyped animals such as proven sires and cows with their own records as reference population to estimate SNP effects. However, the separate steps of estimating SNP effects and EBV during conventional evaluation cannot fully account for genomic preselection; therefore, genomic EBV (GEBV) can be biased (Patry and Ducrocq, 2011). In contrast to the multi-step model, the single-step genomic BLUP (ssGBLUP) model can evaluate genotyped and non-genotyped animals jointly, thus providing unbiased predictive value (Aguilar *et al.*, 2010). The ssGBLUP integrates the genomic relationship matrix (G) and the pedigree relationship matrix (A) into the hybrid matrix and replaces the pedigree relationship matrix in the mixed model equations (MME) of BLUP. On the other hand, Liu *et al.* (2014, 2016) developed a single-step SNP BLUP (ssSNP-BLUP) model equivalent to ssGBLUP in theory. The model does not require the constructions of G and its inversion and can estimate GEBV and SNP effects simultaneously. Liu *et al.* (2014) proposed a computing strategy by separating genomic terms from the MME of ssSNP-BLUP and solving two separate sets of equations. They also showed how to consider reference animals, as defined in the multi-step model by expanding the equations associated with genomic term. However, since the calculation of the two separated equations is repeated alternately, the program may not be easy to implement for complex models.

The objectives of this study were to propose a method that considers a reference population for ssSNP-BLUP model without separating the genomic terms to estimate SNP effects and compare the predictions using different reference populations.

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_311, © T. Osawa *et al.* 2022

Materials & methods

Equation for original ssSNP-BLUP model (ssORG) by Liu *et al.* (2014) can be written as follows,

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}W_1} & \mathbf{X'R^{-1}W_2} & \mathbf{0} \\ \mathbf{W_1'R^{-1}X} & \mathbf{W_1'R^{-1}W_1} + \sigma_u^{-2}\mathbf{\Sigma^{11}} & \mathbf{W_1'R^{-1}W_2} + \sigma_u^{-2}\mathbf{\Sigma^{12}} & \sigma_u^{-2}\mathbf{\Sigma^{1g}} \\ \mathbf{W_2'R^{-1}X} & \mathbf{W_2'R^{-1}W_1} + \sigma_u^{-2}\mathbf{\Sigma^{21}} & \mathbf{W_2'R^{-1}W_2} + \sigma_u^{-2}\mathbf{\Sigma^{22}} & \sigma_u^{-2}\mathbf{\Sigma^{2g}} \\ \mathbf{0} & \sigma_u^{-2}\mathbf{\Sigma^{g1}} & \sigma_u^{-2}\mathbf{\Sigma^{g2}} & \sigma_u^{-2}\mathbf{\Sigma^{3g}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{W_1'R^{-1}y} \\ \mathbf{W_2'R^{-1}y} \\ \mathbf{0} \end{bmatrix} \quad (1)$$

where y is a vector of observations, $\hat{\mathbf{b}}$ is a vector of fixed effects, subscripts 1, 2 and g refer to groups of non-genotyped, genotyped animals and SNP effects, respectively, $\hat{u}_1$ and $\hat{u}_2$ are vectors of additive genetic effects of the non-genotyped and genotyped animals, respectively, $\hat{\mathbf{g}}$ is a vector of SNP effects, X is a design matrix for fixed effects, W_1 and W_2 are design matrices for additive genetic effects of non-genotyped and genotyped animals, respectively, R is $\sigma_e^2\mathbf{I}$, σ_u^2 and σ_e^2 are an additive genetic and residual variance.

$$\Sigma = \begin{bmatrix} \mathbf{\Sigma^{11}} & \mathbf{\Sigma^{12}} & \mathbf{\Sigma^{1g}} \\ \mathbf{\Sigma^{21}} & \mathbf{\Sigma^{22}} & \mathbf{\Sigma^{2g}} \\ \mathbf{\Sigma^{g1}} & \mathbf{\Sigma^{g2}} & \mathbf{\Sigma^{gg}} \end{bmatrix} = \begin{bmatrix} \mathbf{A^{11}} & \mathbf{A^{12}} & \mathbf{0} \\ \mathbf{A^{21}} & \mathbf{A^{22}} + \left(\frac{1}{k} - 1\right)\mathbf{A_{22}^{-1}} & -\frac{1}{k}\mathbf{A_{22}^{-1}Z} \\ \mathbf{0} & -\frac{1}{k}\mathbf{Z'A_{22}^{-1}} & \frac{m}{1-k}\mathbf{I} + \frac{1}{k}\mathbf{Z'A_{22}^{-1}Z} \end{bmatrix} \quad (2)$$

where A^{ij} are submatrices of the inverse of the pedigree relationship matrix, consisting of non-genotyped (1) and genotyped (2), k is the proportion of the additive genetic variance due to the residual polygenic effects (k was set to 0.1 in this analysis), and m is $2\Sigma p_j\,(1 - p_j)$, with p_j being the observed allele frequency of the j^{th} SNP marker and Z is a design matrix of regression coefficients on genotyped animals at all SNP markers ($2\text{-}2p_j$, $1\text{-}2p_j$ or $-2p_j$ for genotype AA, AB, or BB of the j^{th} SNP marker).

Single-Step SNP BLUP model with the reference population. Liu *et al.* (2014) described the genomic term to estimate SNP effects with the reference population as follows.

$$\hat{\mathbf{g}} = \frac{1}{k}\frac{1-k}{m}\mathbf{IZ'FA_{22}^{-1}}\hat{\mathbf{a}}_2 \quad (3)$$

$$\hat{\mathbf{a}}_2 = \hat{\mathbf{u}}_2 - \mathbf{Z}\hat{\mathbf{g}} \quad (4)$$

where F is a filter matrix which is *diag*{1,0,0,1, … 1,1,1,0} to define genotyped animals as reference (=1) or not (=0). Substituting Equation 4 into Equation 3:

$$\hat{\mathbf{g}} = \frac{1}{k}\frac{1-k}{m}\mathbf{IZ'FA_{22}^{-1}}(\hat{\mathbf{u}}_2 - \mathbf{Z}\hat{\mathbf{g}}) \quad (5)$$

Equation 5 can be rearranged as follows:

$$\left[-\frac{1}{k}\mathbf{Z'FA_{22}^{-1}} \quad \frac{m}{1-k}\mathbf{I} + \frac{1}{k}\mathbf{Z'FA_{22}^{-1}Z}\right]\begin{bmatrix}\hat{\mathbf{u}}_2 \\ \hat{\mathbf{g}}\end{bmatrix} = \mathbf{0} \quad (6)$$

Replace $\mathbf{\Sigma^{g2}}$ and $\mathbf{\Sigma^{gg}}$ in Equation 2 by Equation 6 as follows:

$$\mathbf{\Sigma^{g2}} = -\tfrac{1}{k}\mathbf{Z'FA_{22}^{-1}} \text{ and } \mathbf{\Sigma^{gg}} = \tfrac{m}{1-k}\mathbf{I} + \tfrac{1}{k}\mathbf{Z'FA_{22}^{-1}Z} \quad (6)$$

Therefore, the equation of ssSNP-BLUP model with reference population (ssRP) is a nonsymmetric linear system.

Solving algorithm. ssORG is applicable with the preconditioned conjugate gradient method. However, ssRP results in a nonsymmetric linear system, and we adopted the preconditioned biconjugate gradient

stabilized method (Van der Vorst, 1992). The calculation programs introduced an 'iteration on data' algorithm (Strandén and Lidauer., 1999 and Tsuruta *et al.*, 2001) and used 10 OpenMP threads. The preconditioner matrix combined the Jacobi preconditioner and a second-level diagonal preconditioner (Vandenplas *et al.*, 2019). Assuming that Equation 1 is written as Cx = b, the convergence criterion is defined as $\|b - Cx\|^2 / \|b\|^2 < 10^{-14}$, which $\|\cdot\|$ is the 2-norm. In addition, the average of the last 30 iteration rounds must be less than 10^{-14} in the biconjugate gradient stabilized method.

Data. We used overall conformation score in the first lactation for Japanese Holsteins. The additive genetic variance, residual variance and heritability of the trait were 0.685, 1.852 and 0.27, respectively. The dataset comprised 762,650 cows with record and 8,745 sires with daughters; of those, 37,197 cows and 5,352 sires were genotyped. The dataset also included 88,058 genotyped heifers and 3,973 genotyped young bulls. After the quality control, 39,756 SNP markers were considered in this study. We used three reference populations: sires with daughters (ssRP_S), cows with the record (ssRP_C) and ssRP_SC with both of them.

Results

In ssORG, 821 iterations were needed to stratify convergency criteria, and the total computing time was 94 minutes. For ssRP_S, ssRP_C and ssRP_SC, although the number of iterations was less than ssORG (424 to 600), the total times was slightly longer (106 to 140 minutes).

Correlations and regression coefficients. Pearson correlations in the solutions between ssORG and ssRP and regression coefficients of ssORG on ssRP for SNP effects and GEBV for genotyped sires, cows, young bulls and heifers are showed in Table 1. Correlations in ssRP_SC were closed to 1.0, but ssRP_S and ssRP_C were lower than ssRP_SC. Therefore, although the contribution of young animals to the predictions was small, the composition of reference population affected the predicted values. Since the regression coefficients were smaller than 1.0 for almost any combinations, considering reference population may reduce the overestimation of the predictions.

Genetic trends of genotyped animal. Genetic trends of genotyped sires and young bulls for ssORG and ssRP are shown in Figure 1. Similar trends were obtained for both sires and young bulls in ssORG, ssRP_S and ssRP_SC. However, the genetic trends of ssRP_C were flattened in older generations of sires and young bulls, compared to other reference populations.

Table 1. Correlations between ssORG[1] and ssRP[2] and regression coefficients of ssORG on ssRP for SNP effects and GEBV for genotyped sire, cows, young bulls and heifers.

		Correlations			Regression coefficients		
Items	**n**	**ssRP_S**	**ssRP_C**	**ssRP_SC**	**ssRP_S**	**ssRP_C**	**ssRP_SC**
SNP effects	39,756	0.531	0.790	0.967	0.384	0.740	0.950
Sires	5,332	0.979	0.975	0.999	0.979	0.881	1.001
Cows	37,197	0.894	0.962	0.995	0.838	0.931	0.987
Young bulls	3,973	0.865	0.915	0.992	0.852	0.818	0.985
Heifers	88,058	0.880	0.936	0.993	0.838	0.883	0.972

[1] Original single-step SNP BLUP model.

[2] Single-step SNP BLUP model with reference populations (S=sires, C=cows and SC=sires and cows).

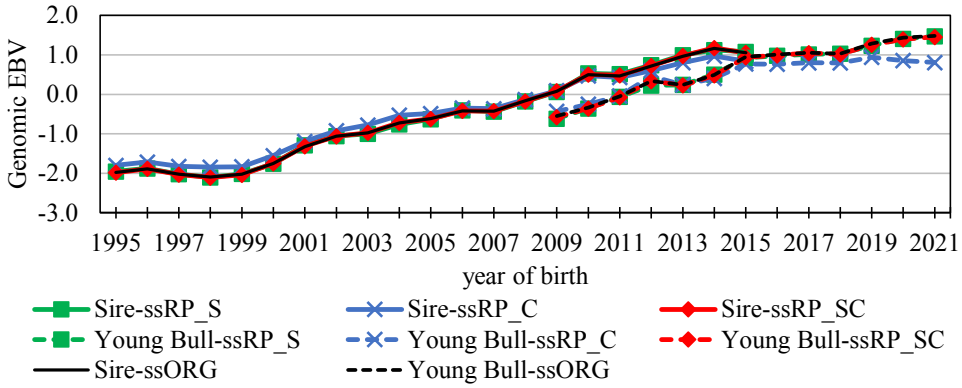

Figure 1. Genetic trends of genotyped sires and young bulls by models. ssORG = Original single-step SNP BLUP model, ssRP = Single-step SNP BLUP model with reference populations (S=sires, C=cows and SC=sires and cows).

Discussion

The proposed method can consider the reference animals without separating the genomic term from the MME in ssSNP-BLUP. This model is easily extensible to complex models e.g. multi-trait and random regression models (results is not shown). In addition, this method can be helpful to understand the influence of the reference population on GEBV and SNP effects in the single-step method. In the results of this analysis, although the contribution of young animals to both GEBV and SNP effects was small, the predictions varied greatly when only sires or cows were considered as reference population. Thus, the composition of the reference population may also need to be carefully considered in the single-step methods. Considering the reference population may reduce the overestimation of the predictions. However, the genetic trends were flattened when only cows were considered in reference population. This may be because the fact that genotyped cows were only relatively new generations. Further studies are needed to investigate the prediction accuracy using the method.

References

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S. and Lawlor T.J. (2010) J. Dairy Sci. 93:743–752. https://doi.org/10.3168/jds.2009-2730

Liu, Z., Goddard M.E., Hayes B.J., Reinhardt F., and F., and Reents R. (2016) J. Dairy Sci. 99:2016-2025. https://doi.org/10.3168/jds.2015-10394

Liu, Z., Goddard M.E., Reinhardt F., and Reents R. (2014) J. Dairy Sci. 97:5833-5850. https://doi.org/10.3168/jds.2014-7924

Patry C., and Ducrocq V. (2011) J. Dairy Sci. 94:1011–1020. https://doi.org/10.3168/jds.2010-3804

Strandén, I., and Lidauer M. (1999) J. Dairy Sci. 82:2779-2787. https://doi.org/10.3168/jds.S0022-0302(99)75535-9

Tsuruta, S., Misztal I., and Strandén I. (2001) J. Anim. Sci. 79:1166-1172. https://doi.org/10.2527/2001.7951166x

Van der Vorst H. (1992) SIAM J. Sci. and Stat. Comput. 13:631–644. https://doi.org/10.1137/0913035

Vandenplas J., Calus M.P.L., Eding H., and Vulik C. (2019) Genet. Sel. Evol. 51:30 https://doi.org/10.1186/s12711-019-0472-8

VanRaden, P. M, 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91:4414-4423. https://doi.org/10.3168/jds.2007-0980