J. Dairy Sci. 102:10012–10019
https://doi.org/10.3168/jds.2019-16262

Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants

B. O. Fragomeni,[1]* D. A. L. Lourenco,[2] A. Legarra,[3] P. M. VanRaden,[4] and I. Misztal[2]
[1]Department of Animal Science, University of Connecticut, Storrs-Mansfield 06269
[2]Department of Animal and Dairy Science, University of Georgia, Athens 30602
[3]Institut National de la Recherche Agronomique, UMR1388 GenPhySE, Castanet Tolosan, France 31326
[4]Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705

ABSTRACT

Causal variants inferred from sequence data analysis are expected to increase accuracy of genomic selection. In this work we evaluated the gain in reliability of genomic predictions, for stature in US Holsteins, when adding selected sequence variants to a pre-existent SNP chip. Two prediction methods were tested: de-regressed proofs assuming heterogeneous (genomic BLUP; GBLUP) residual variances and by single-step GBLUP (ssGBLUP) using actual phenotypes. Phenotypic data included 3,999,631 records for stature on 3,027,304 Holstein cows. Genotypes on 54,087 SNP markers (54k) were available for 26,877 bulls. Additionally, 16,648 selected sequence variants were combined with the 54k markers, for a total of 70,735 (70k) markers. In all methods, SNP in the genomic relationship matrix (**G**) were unweighted or weighted iteratively, with weights derived either by SNP effects squared or by a nonlinear method that resembles BayesA (nonlinear A). Reliability of genomic predictions were obtained by cross validation. With unweighted **G** derived from 54k markers, the reliabilities (× 100) were 72.4 for GBLUP and 75.3 for ssGBLUP. With unweighted **G** derived from 70k markers, the reliabilities were 73.4 and 76.0, respectively. Weighting by nonlinear A changed reliabilities to 73.3, and 75.9, respectively. Addition of selected sequence variants had a small effect on reliabilities. Weighting by quadratic functions reduced reliabilities. Weighting by nonlinear A increased reliabilities for GBLUP but had only a small effect in ssGBLUP. Reliabilities for direct genomic values extracted from ssGBLUP using unweighted **G** with 54k were higher than reliabilities by any GBLUP. Thus, ssGBLUP seems to capture more information than GBLUP and there is less room for extra reliability. Improvements in GBLUP may be because the weights in **G** change the covariance structure, which can explain a proportion of the variance that is accounted for when a heterogeneous residual variance is assumed by considering a different number of daughters per bull.

Key words: causative variant, BayesA, genomic prediction, sequence data

INTRODUCTION

Genomic selection uses dense SNP, not genes or QTL per se. An increase of up to 4 percentage points in reliability, in particular across generations or populations, can occur if causative loci are identified and given appropriate weights (Brondum et al., 2015), so that there are no linkage disequilibrium phase changes and the effect of the quantitative trait loci markers are not excessively regressed toward 0. If all causative SNP are identified and their substitution effect is constant, genomic selection should approach 100% accuracy.

In dairy cattle, complete sequence data were generated by the 1000 bulls genome project (Hayes et al., 2014). Several studies used SNP subsets from sequence data aiming to improve genomic predictions. Although some studies showed no improvement compared with using a regular SNP chip (Veerkamp et al., 2016), up to a 5% increase in accuracy was achieved by VanRaden et al. (2017), when applying nonlinear A (similar to BayesA) to a set of 60k "regular" SNP versus a subset of the 60k plus 17k potentially causative SNP from sequence polymorphisms that had been identified by GWAS across traits.

Single-step GBLUP (**ssGBLUP**) is a popular and flexible method for genomic predictions. Fragomeni et al. (2017) investigated the possibility of utilizing causative variants in ssGBLUP in a simulation study, where a "weighted" genomic relationship matrix (**G**), as de-

Received January 7, 2019.
Accepted July 16, 2019.
*Corresponding author: breno.fragomeni@uconn.edu

fined in Wang et al. (2012), was used to place more emphasis to particular markers. With unweighted **G**, the inclusion of causative SNP increased accuracy by a maximum of 0.04. After weighting causative variants, accuracy increased from 0.07 to 0.30 depending on the strategy used (Fragomeni et al., 2017).

The SNP weights commonly used in weighted ssGBLUP (Wang et al., 2012) are calculated based on SNP solution squared. Some of these methods, when used iteratively, diverge as large effects become too large and some become too small. Zhang et al. (2016) looked into alternative methods for SNP weighting. Nonlinear A (VanRaden, 2008) reduces and limits changes in SNP weight, converges, and is routinely used in dairy cattle. Other methods [quadratic (Wang et al., 2012), FastBayesA (Sun et al., 2012)] have rarely been used in real data and diverge in many cases.

In addition, the increases in accuracy by using putative causal variants from sequence data in VanRaden et al., (2017) were obtained using daughter yield deviations (i.e., transformed data from regular pedigree-based evaluations). The effect of this strategy, in real data, when using a more comprehensive method such as ssGBLUP is still unknown.

The purpose of this study is to evaluate accuracy of GBLUP and ssGBLUP, for stature in US Holsteins, using regular SNP and selected sequence variants with different methods to calculate SNP weights.

MATERIALS AND METHODS

Field Data

Among the 33 traits used for US Holsteins genetic evaluation, stature was the one chosen for this study. This is because the gain in reliability reported by VanRaden et al. (2017) after including selected sequence variants to an existing SNP chip was the greatest for this specific trait. Phenotypic data were provided by Holstein Association USA Inc. (Brattleboro, VT). The data were restricted to phenotypes recorded between 1990 and 2016, and pedigrees were traced 3 generations back. The data included 3,999,631 records for stature on 3,027,304 cows. Pedigree information was available for 4,661,872 animals. Genotypes on 54,087 SNP markers were available for 26,877 bulls. Additionally, in a separate analysis, a total of 16,648 selected sequence variants (VanRaden et al., 2017) were included in the genotype file, totaling 70,735 markers. Those variants were selected based on absolute effect size for 33 traits evaluated in US Holsteins. As explained in VanRaden et al., (2017), 1000 sequence variants were selected for each trait, but after removing duplicates only 16,648 variants remained. More information about how the selected sequence variants were chosen can be found in VanRaden et al. (2017).

In our study we investigated only one trait but sequence variants obtained based on 33 traits were used. This is because using a pooled set of variants instead of trait-specific variants avoids the need of having a SNP chip/file for each trait and is the logical choice for multivariate systems (i.e., as the official evaluations). Additionally, using the same set of selected variants as in VanRaden et al. (2017) allows for a better comparison of methods.

For validation purposes, a reduced data set was constructed that consisted of phenotypes up to 2011. A total of 2,521 validation bulls were born before 2010 and had no daughters with records in the reduced data set, but at least 30 daughters in the complete data set.

Methods

Both GBLUP and ssGBLUP were used in this study. For GBLUP, only genotyped animals were used for evaluation. If we consider **a** to be a vector of additive genetic effects, the distribution of **a** under GBLUP is as follows:

$$\mathbf{a} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_a^2\right),$$

where **G** is a genomic relationship matrix, and σ_a^2 is the additive genetic variance. When only a fraction of animals is genotyped, single-step GBLUP (**ssGBLUP**) combines pedigree and genomic information into a realized relationship matrix (**H**), such that $\mathbf{a} \sim N\left(\mathbf{0}, \mathbf{H}\sigma_a^2\right)$. Its inverse has a simple form (Aguilar et al., 2010), where **A** is the numerator relationship matrix:

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

The genomic relationship matrix was constructed based on VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{M}\mathbf{M}'}{2\sum_i^m p_i q_i},$$

where **M** is a centered matrix of SNP content, p_i and q_i are gene frequencies, and m is the number of SNP. This approach assumes all SNP effects (u) follow a multivariate normal distribution, which may not be biologically true because causative variants are expected to have large effects. Markers that are causal or in linkage

disequilibrium with causative variants should be given higher weights. To account for heterogeneous SNP weights, a matrix of weights should be included in the formula for constructing **G**, where $var(\mathbf{s})$ is the vector containing the variance of individual SNP effects, and d_i is the ith diagonal element of **D**, accounting for the ith SNP weight:

$$var(\mathbf{s}) = \mathbf{D} = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_m \end{bmatrix}.$$

Based on that, a weighted relationship matrix can be defined as

$$\mathbf{G_w} = \frac{\mathbf{MDM'}}{\sum_{j=1}^{m} 2p_j q_j},$$

where **D** is a matrix of weights and each diagonal element of this matrix is defined as

$$d_i = \sigma^2_{u,i} \frac{\sum_{j=1}^{m} 2p_j q_j}{\sigma^2_a},$$

where $\sigma^2_{u,i}$ can be understood as SNP "prior variances" (Meuwissen et al., 2001; Gianola, 2013). In practice, $\sigma^2_{u,i}$ are not known (or even well defined; Gianola, 2013). Several approaches exist to estimate individual SNP variances, but $\sigma^2_{u,i}$ can be approximated from estimates of SNP effects as follows. Once genomic predictions are obtained by GBLUP or ssGBLUP, SNP effects can be calculated using a backsolving process (Stranden and Garrick, 2009; VanRaden, 2008; Wang et al., 2012):

$$\hat{\mathbf{u}} = \frac{1}{\sum_{j=1}^{m} 2p_j q_j} \mathbf{DM'G_w^{-1}\hat{a}},$$

where $\hat{\mathbf{u}}$ is the vector of estimated SNP effects, and $\hat{\mathbf{a}}$ is a vector of genomic EBV (**GEBV**). Once SNP effects are calculated, variances $\sigma^2_{u,i}$ can be estimated in 2 ways:

Quadratic weights: $\sigma^2_{u,i} = \hat{u}_i^2$. This is the marker weighting methodology used by Fragomeni et al. (2017), and differs from the originally proposed weights ($\hat{u}_i^2 2pq$; Wang et al., 2012) because allele frequencies are not considered.

Nonlinear A: $\sigma^2_{u,i} = \frac{\sigma^2_a}{\sum_{j=1}^{m} 2p_j q_j} 1.25^{\frac{|\hat{u}_i|}{sd(\hat{u})}-2}$. This method was described by VanRaden (2008) and Cole et al. (2009) and is a faster, analogous version of BayesA. In the nonlinear A formula, $|\hat{u}_i|$ is the absolute estimated SNP effect for marker i, and $sd(\hat{u})$ is the standard deviation of the vector of estimated SNP effects. The parameter value of 1.25, proposed by VanRaden (2008) and 1.2 by Cole et al. (2009), is a constant that determines how much the distribution of SNP effects departs from the normal distribution (i.e., 1 means normal distribution). Constant values from 1.115 to 1.40 were also tested. The maximum change in weights was limited by capping the $\frac{|\hat{u}_i|}{sd(\hat{u})}$ term at 5 or 10 to avoid extreme SNP effects and to improve convergence. This algorithm is regularly used for national US Holstein genomic evaluations.

All analyses were run iteratively, updating GEBV and SNP weights for 10 iterations. The GEBV were calculated in every iteration, and convergence was achieved when changes in those values, from current to previous iteration, were less than 10^{-4}. Additionally, convergence was assessed by visual inspection of reliability plots. Analysis with quadratic weights converged at very low reliability values; therefore, results are shown for the second iteration. Reliabilities with the quadratic method are expected to decrease after a few iterations; consequently, it is recommended to limit the number of iterations for maximum predictivity (Zhang et al., 2016). Fast BayesA was not attempted in this study as initial simulations showed that although it was more reliable than the quadratic method, nonlinear A converged more consistently (Fragomeni et al., 2018).

Quadratic weights and nonlinear A were compared with a benchmark scenario where all SNP were assumed to explain the same proportion of additive genetic variance. This scenario was called unweighted (ss)GBLUP.

Analyses

Data were analyzed with GBLUP and ssGBLUP, using pseudo-phenotypes for the first and regular phenotypes with a more comprehensive model for the last. Pseudo-phenotypes were computed as deregressed evaluations (hereby termed DD to avoid confusion with the daughter equivalent) following formulas described in VanRaden et al. (2009) and Wiggans et al. (2011); the reduced data set with phenotypes up to 2011 was used:

$$DD = PA + \left[\frac{EBV - PA}{\frac{DE_{Prog}}{(DE_{Prog} + DE_{PA} + 1)}}\right],$$

where the term $\frac{DE_{Prog}}{(DE_{Prog} + DE_{PA} + 1)}$ is a measure of reliability of DD (R_{DD}) and is shown in Wiggans et al. (2011)

$$DE_{PA} = \frac{REL_{PA}}{(1 - REL_{PA})},$$

$$DE_{Prog} = \left[\frac{REL_{EBV}}{(1 - REL_{EBV})}\right] - DE_{PA},$$

where PA is parent average, DE_{PA} is daughter equivalent from parent average, DE_{Prog} is daughter equivalent from progeny information, REL_{PA} is reliability of PA, and REL_{EBV} is reliability of EBV. Daughter equivalent formulas from VanRaden and Wiggans (1991) include the variance ratio k; however, it is simplified here by dividing the numerator and denominator by k to make DE unitless and to allow simpler comparison across traits. The average (SD) DE_{Prog} for validation animals based on complete data was 25 (101).

The model used for GBLUP is as follows:

$$\mathbf{y_p} = \mathbf{1}b + \mathbf{Za} + \mathbf{e},$$

where $\mathbf{y_p}$ is the vector of DD, b is the general mean, $\mathbf{a}$ is the additive genetic effect, $\mathbf{e}$ is the residual effect, and $\mathbf{Z}$ is the incidence matrix for the effects contained in $\mathbf{a}$. Two different distributions were considered for the residuals; a crude one of homogeneous residual variance, where $e \sim N\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$, and $\mathbf{I}$ is the identity matrix, σ_e^2 is the residual variance, and a more refined one of heterogeneous residual variance with $e \sim N\left(\mathbf{0}, \mathbf{R}\sigma_e^2\right)$, with $\mathbf{R}$ being a diagonal matrix with elements equal to the reciprocal of daughter equivalent for each genotyped bull. This accounts for the fact that bulls have different amounts of information used for computing DD. The homogeneous residual variance approach is not acceptable in commercial implementations of GBLUP because of its inappropriate assumption of equal contribution for all sires. It was still considered in the present study to compare the effect of SNP weights and inclusion of selected variants in a less optimal model.

As no polygenic effect was added to GBLUP, we are going to refer to the predictions obtained in this model as direct genomic values (**DGV**).

For ssGBLUP, the model was

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{bf} + \mathbf{W}\boldsymbol{hs} + \mathbf{Z}\boldsymbol{a} + \boldsymbol{e},$$

with $\boldsymbol{bf}$ containing fixed effects of herd-year-season, age-parity, and lactation stage-parity, and $\boldsymbol{hs}$ containing the random effect of herd-sire; $\mathbf{X}$ is the incidence matrix for the effects contained in $\boldsymbol{bf}$; and $\mathbf{W}$ is the incidence matrix for the effects contained in $\boldsymbol{hs}$. More details about the model used for stature are described in Tsuruta et al. (2002).

Model Validation

Adjusted reliability for validation bulls when using genomic models (REL_{GEN}) was calculated based on the reduced data set, following the procedure described by VanRaden et al. (2009):

$$REL_{GEN} = \frac{R2_{GEN}}{R_{DD}} + (REL_S + REL_D)/4 - \left(\frac{R2_{PA}}{R_{DD}}\right),$$

where REL_S and REL_D are the EBV reliabilities of sire and dam, respectively. The $R2_{PA}$ is the reliability of PA calculated as the coefficient of determination of the regression of daughter deviation in the complete data set (DD_{2016}) on PA in the reduced data set (PA_{2011}). This regression was weighted by the reliability of DD in the complete data set (R_{DD_2016}), and R_{DD} is the average value of R_{DD_2016}. The model fitted was

$$DD_{2016} = b0 + b1 \times PA_{2011},$$

where DD_{2016} was calculated using formulas previously shown; b0 is the intercept and b1 is the slope. The latter was used as a measure of inflation. The $R2_{GEN}$ is the coefficient of determination of genomic predictions, which was calculated using the above regression formula after replacing PA_{2011} by DGV_{2011} (i.e., for GBLUP) or $GEBV_{2011}$ (i.e., for ssGBLUP). As a final result, we provided adjusted reliabilities to make our results comparable to the ones in VanRaden et al. (2017).

When GBLUP is used, the genomic prediction is only based on genomic information and progeny contribution (i.e., in the form of DD), which makes comparisons between predictions based on GBLUP and ssGBLUP unfair. This is because ssGBLUP also accounts for pedigree information and uses raw phenotypes. In this way, for validation animals, we also computed DGV

from ssGBLUP as the sum of SNP effects weighted by the SNP content.

RESULTS AND DISCUSSION

Adjusted reliabilities rather than coefficients of determination are shown to ease comparisons between our results and the ones presented in VanRaden et al. (2017). The adjustments convert the squared correlations into published genomic reliabilities. On average, adjustments added 14.2 points to coefficients of determination. Adjusted reliabilities of genomic predictions from GBLUP and ssGBLUP without weights, with quadratic and nonlinear A weights are in Table 1. The adjusted reliabilities reported for quadratic weight were from the second iteration, which had the highest value over all iterations. Although the visual check is not optimal, it is a common practice when quadratic weights are used because of lack of a convergence criteria (Wang et al., 2012, Zhang et al., 2016). Conversely, nonlinear A converges easily, and in our study the convergence was reached at the highest accuracy.

When using GBLUP with heterogeneous residual variance and quadratic weights for SNP, reliabilities decreased by 2.2 and 2.9 points for 54k and 70k (i.e., 54k + 17k causative variants), compared with the unweighted approach. No difference was observed when using the nonlinear A rather than the unweighted approach. However, adding the 17k selected variants did increase reliability in the GBLUP with homogeneous residual variance. This increase was stable with nonlinear A but it deteriorated after the second iteration with quadratic weights. However, even after weighting, the model assuming homogeneous residual variance did not reach the level of accuracy of the 2 more robust models: GBLUP with heterogeneous residual variance and ssGBLUP.

For ssGBLUP, adding selected variants with a proper weight only slightly increased reliabilities, however the increment was less than the observed by VanRaden et al. (2017). Similar results were found without adding weights (Table 1). Additionally, we applied the SNP weights calculated for stature by VanRaden et al. (2017), but again, reliability was 76.0. Although no improvements were observed whether weights were used or not, reliability from ssGBLUP was 2.7 points greater than the best GBLUP heterogeneous scenario (73.4). One possible explanation for the discrepancy in results between methods is that ssGBLUP deals with more information than multistep procedures, especially when

Table 1. Reliability for genomic BLUP (GBLUP) and single-step GBLUP (ssGBLUP) when including selected variants to the 54k SNP panel under different weighting approaches

Method	Adjusted reliability	Gain from adjusted reliability compared with parent average	Gain from adjusted reliability compared with DGV[1] from ssGBLUP
Parent average	38.6	0.0	0.0
GBLUP, homogeneous[2]			
Unweighted 54k	68.8	30.2	−5.5
Unweighted 70k	69.5	30.9	−4.8
Quadratic 54k	68.4	29.8	−5.9
Quadratic 70k	68.3	29.7	−6.0
Nonlinear A 54k	70.8	32.2	−3.5
Nonlinear A 70k	70.9	32.3	−3.4
GBLUP, heterogeneous[3]			
Unweighted 54k	72.4	33.8	−1.9
Unweighted 70k	73.4	34.7	−1.0
Quadratic 54k	70.5	31.9	−3.8
Quadratic 70k	70.4	31.8	−3.9
Nonlinear A 54k	73.2	34.6	−1.1
Nonlinear A 70k	73.3	34.7	−1.0
ssGBLUP, GEBV[4]			
Unweighted 54k	75.3	36.7	1.0
Unweighted 70k	76.0	37.4	1.7
Quadratic 54k	71.6	33.0	−2.7
Quadratic 70k	72.1	33.5	−2.2
Nonlinear A 54k	75.5	36.9	1.2
Nonlinear A 70k	75.9	37.3	1.6
ssGBLUP, DGV[5]			
Unweighted 54k	74.3	35.7	0.0

[1]DGV = direct genomic value.

[2]GBLUP considering homogeneous residual variance.

[3]GBLUP considering heterogeneous residual variance weighted by daughter equivalent.

[4]GEBV obtained by ssGBLUP.

[5]DGV obtained by SNP effects from ssGBLUP.

de-regressions are used in the latter. In this way, any assumption made a priori about the SNP effects is overwhelmed by the data in ssGBLUP, whereas in GBLUP each extra piece of information may have some effect. Another reason is the explicit contribution of parent average in ssGBLUP. In addition, some reliability can be lost in multistep because of approximations used in the deregression step.

Karaman et al. (2016) showed that accuracies from weighted and unweighted multistep methods converged to the same value as the amount of data increased (i.e., number of genotyped animals). Lourenco et al. (2017) used simulations to demonstrate that SNP weighting is not needed in ssGBLUP when the genotyped population is large (>5k). This was true for oligogenic and polygenic traits. For the latter, weighting harmed predictions even when the number of genotyped animals was as low as 2k. For oligogenic traits or when few SNP show strong effect, as in the case of milk contents, strongly affected by DGAT1 (Grisart et al., 2004), SNP weighting has shown to be beneficial in small data sets, although its benefit decreases for large data sets (VanRaden et al., 2009, 2011). In fact, quadratic weights increased prediction accuracy for bacterial cold water disease in rainbow trout (Vallejo et al., 2016) and fat and protein percentages in a small Israeli Holstein population (Lourenco et al., 2014) using weighted ssGBLUP, compared with unweighted ssGBLUP.

Figure 1 shows adjusted reliabilities for all 10 iterations when the selected variants were present in the data. Iteration 1 means equal weights were assigned for all SNP. If a plateau is reached, the convergence is obtained and SNP variances do not change after that, keeping the reliability steady. It is clear that quadratic weights in GBLUP and ssGBLUP tend to diverge. In fact, the same result was also observed with simulated data (results not shown). Drops in reliability were not observed for the nonlinear A weights.

The extent of inflation (b1) in the estimates depended on the method; however, weights also affected b1 in a smaller scale (Table 2). The GBLUP methods had b1 values between 0.88 and 0.9 when **G** was unweighted or weighted by the nonlinear A method. Quadratic weights resulted in b1 lower than 0.7. Under ssGBLUP, b1 values were 0.87 or 0.88 for unweighted **G** and nonlinear A, whereas quadratic weights had b1 values as low as 0.79. The DGV predictions with ssGBLUP lead to the least inflated values. Those results show that multi-step procedures return less inflated estimates.

When large causative variants are included in the model, we expect an increase in reliability by properly accounting for them. Among the weights we tested, it was clear that the quadratic approach was unable to correctly use the selected variants, because of the extreme changes in weights at each iteration. Nonlinear A method is suitable for polygenic traits; however, its implementation for weight updates in ssGBLUP may not have resulted in maximum reliability. Overall, the lack of improvement in reliability for ssGBLUP can be because the selected variants may have had only small effects, and because given the large data set, the prior information on SNP effects does not change the final result. The increase in accuracy using weights in homogeneous GBLUP suggests that weights somehow compensate for model weaknesses. The exact mechanism is not clear, but a possible hypothesis is a better model fitting because of extra flexibility provided by weights. Therefore, different SNP weights in **G** could be implicitly explaining some of the variance due to different number of daughters per bull.

This study raises several questions. Was any improvement in VanRaden et al. (2017) due to problems

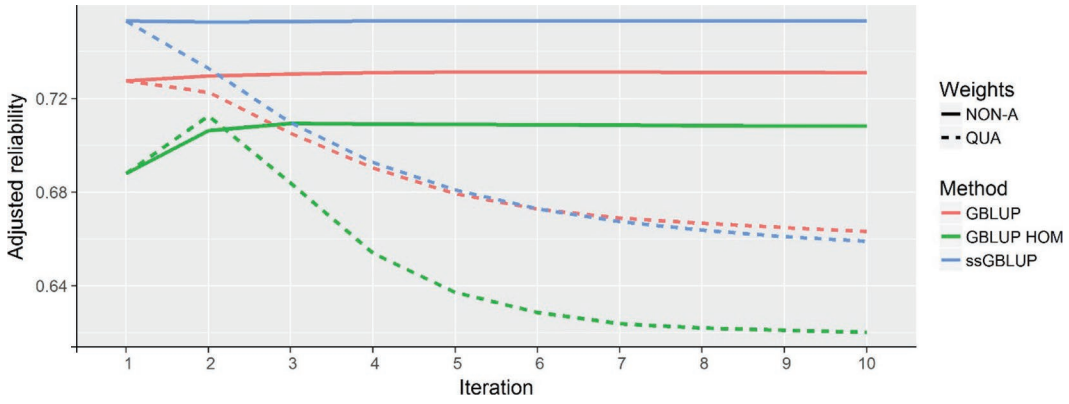

Figure 1. Adjusted reliabilities from direct genomic value obtained by genomic BLUP (GBLUP) and genomic EBV (GEBV) obtained by ssGBLUP with weighted genomic relationship matrix in the first 10 iterations. The GBLUP assumed heterogeneous residual variance (GBLUP) and homogeneous residual variance (GBLUP HOM). Weights used were calculated by the nonlinear A (NON-A) and quadratic weights (QUA) when the causative variants were in the data (i.e., 70k SNP markers).

Table 2. Coefficient from regression of daughter deviation on genomic prediction (b1)[1]

Item	b1	Limit	Constant
Parent average	0.82		
GBLUP, homogeneous[2]			
Unweighted 54k	1.01		
Unweighted 70k	1.00		
Linear 54k	0.70		
Linear 70k	0.70		
Nonlinear A 54k	0.96	5	1.400
Nonlinear A 70k	0.94	5	1.400
GBLUP, heterogeneous[3]			
Unweighted 54k	0.90		
Unweighted 70k	0.89		
Linear 54k	0.67		
Linear 70k	0.68		
Nonlinear A 54k	0.88	10	1.175
Nonlinear A 70k	0.88	5	1.115
ssGBLUP, GEBV[4]			
Unweighted 54k	0.88		
Unweighted 70k	0.88		
Linear 54k	0.79		
Linear 70k	0.79		
Nonlinear A 54k	0.87	10	1.200
Nonlinear A 70k	0.88	5	1.125
ssGBLUP, DGV[5]			
Unweighted 54k	0.92		

[1]When the nonlinear A formula $\sigma^2_{u,i} = \frac{\sigma^2_a}{\sum_{j=1}^{m} 2p_j q_j} c^{\frac{|\hat{u}_i|}{sd(\hat{u})}-2}$ was used, the constant was the value attributed to c, and limit was the maximum value allowed to $\frac{|\hat{u}_i|}{sd(\hat{u})}$, where $\sigma^2_{u,i}$ is the SNP prior variance for the ith SNP, σ^2_a is the additive genetic variance, p_j and q_j are the allelic frequencies for the jth SNP, m is the number of SNP, $\hat{u}_i$ is the ith SNP effect estimate, $sd(\hat{u})$ is the sd of the SNP effects estimates, and c is a constant.

[2]GBLUP considering homogeneous residual variance.

[3]GBLUP considering heterogeneous residual variance weighted by daughter equivalent.

[4]Genomic EBV (GEBV) obtained by ssGBLUP.

[5]Direct genomic value (DGV) obtained by SNP effects from ssGBLUP.

with pseudo-observations? However, similar improvements were obtained by VanRaden et al. (2017) using simulated data, where these problems were absent. Is ssGBLUP, which uses all information, less sensitive to SNP weighting? In a previous study, assigning simulated true weights to causative loci in ssGBLUP was effective in improving reliabilities (Fragomeni et al., 2017), but the number of causative loci was reduced. Also, there is an implicit assumption that causal loci are biallelic, which does not necessarily hold.

Using trait-specific **G** for many traits in ssGBLUP may be expensive and may prohibit multiple-trait models. On the other hand, using no weights provides simple implementation. For instance, genomic evaluation in Angus includes multiple traits with maternal effects and uses ssGBLUP with unweighted **G** for regular weekly evaluations, and SNP predictions derived from GEBV of ssGBLUP for rapid interim predictions (Lourenco et al., 2015).

CONCLUSIONS

For stature, improvements in GBLUP may be because the weights in **G** change the covariance structure, which can explain a proportion of the variance that is accounted for when weighting the observations by the number of daughters per bull (i.e., using heterogeneous residual variance). Using quadratic weights for SNP is not beneficial when selected variants have a small effect, because of extreme values. Nonlinear A is stable and more accurate than quadratic weights, and its use is highly advised. Adding selected variants slightly increases reliabilities for single-step GBLUP. Therefore, this method seems less affected by prior information about the genetic architecture of the trait and provides greater accuracy/reliability than multistep methods.

ACKNOWLEDGMENTS

This study was partially funded by Agriculture and Food Research Initiative Competitive Grants no. 2015-67015-22936 from the US Department of Agriculture's National Institute of Food and Agriculture. The authors thank Mel Tooker (Animal Genomics and Improvement Laboratory, Beltsville, MD) for assistance with data preparation, Duane Norman (Council on Dairy Cattle Breeding, Bowie, MD) for helpful contributions to the manuscript, and the anonymous reviewers for their comments and suggestions. The contribution of dairy producers who supplied data through their participation in the Dairy Herd Improvement program and the Dairy Records Processing Centers that edited and relayed information on to the Council of Dairy Cattle Breeding are also acknowledged.

REFERENCES

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752.

Brondum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J. Dairy Sci. 98:4107–4116. https://doi.org/10.3168/jds.2014-9005.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92:2931–2946. https://doi.org/10.3168/jds.2008-1762.

Fragomeni, B. O., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2017. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genet. Sel. Evol. 49:59. https://doi.org/10.1186/s12711-017-0335-0.

Fragomeni, B. O., D. A. L. Lourenco, M. E. Tooker, P. M. Van Raden, and I. Misztal. 2018. Use of causative variants and SNP weighting in a single-step GBLUP context. In Proc. 11th World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand.

Gianola, D. 2013. Priors in whole-genome regression: The Bayesian alphabet returns. Genetics 194:573–596. https://doi.org/10.1534/genetics.113.151753.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim, A. Kvasz, M. Mni, P. Simon, J. M. Frere, W. Coppieters, and M. Georges. 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. USA 101:2398–2403. https://doi.org/10.1073/pnas.0308518100.

Hayes, B. J., I. M. MacLeod, and H. D. Daetwyler. B. J. Phil, A. J. Chamberlain, C. Vander Jagt, A. Capitan, H. Pausch, P. Stothard, X. Liao, and C. Schrooten. 2014. Genomic prediction from whole genome sequence in livestock: The 1000 bull genomes project. Communication 183 in Proc. 10th World Congress on Genetics Applied to Livestock Production. Vancouver, Canada. Am. Soc. Anim. Sci., Champaign, IL.

Karaman, E., H. Cheng, M. Z. Firat, D. J. Garrick, and R. L. Fernando. 2016. An upper bound for accuracy of prediction using GBLUP. PLoS One 11:e0161054. https://doi.org/10.1371/journal.pone.0161054.

Lourenco, D. A. L., B. O. Fragomeni, H. L. Bradford, I. R. Menezes, J. B. S. Ferraz, I. Aguilar, S. Tsuruta, and I. Misztal. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. J. Anim. Breed. Genet. 134:463–471. https://doi.org/10.1111/jbg.12288.

Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, E. Ezra, M. Ron, A. Shirak, and J. I. Weller. 2014. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. J. Dairy Sci. 97:1742–1752. https://doi.org/10.3168/jds.2013-6916.

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and I. Misztal. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93:2653–2662. https://doi.org/10.2527/jas.2014-8836.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Stranden, I., and D. J. Garrick. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92:2971–2975. https://doi.org/10.3168/jds.2008-1929.

Sun, X., L. Qu, D. J. Garrick, J. C. Dekkers, and R. L. Fernando. 2012. A fast EM algorithm for BayesA-like prediction of genomic breeding values. PLoS One 7:e49157. https://doi.org/10.1371/journal.pone.0049157.

Tsuruta, S., I. Misztal, L. Klei, and T. J. Lawlor. 2002. Analysis of age-specific predicted transmitting abilities for final scores in Holsteins with a random regression model. J. Dairy Sci. 85:1324–1330. https://doi.org/10.3168/jds.S0022-0302(02)74197-0.

Vallejo, R. L., T. D. Leeds, B. O. Fragomeni, G. Gao, A. G. Hernandez, I. Misztal, T. J. Welch, G. D. Wiens, and Y. Palti. 2016. Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in rainbow trout: Insights on genotyping methods and genomic prediction models. Front. Genet. 7:96. https://doi.org/10.3389/fgene.2016.00096.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43:10. https://doi.org/10.1186/1297-9686-43-10.

VanRaden, P. M., M. E. Tooker, J. R. O'Connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet. Sel. Evol. 49:32. https://doi.org/10.1186/s12711-017-0307-4.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24. https://doi.org/10.3168/jds.2008-1514.

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74:2737–2746.

Veerkamp, R. F., A. C. Bouwman, C. Schrooten, and M. P. Calus. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. Genet. Sel. Evol. 48:95. https://doi.org/10.1186/s12711-016-0274-1.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. (Camb.) 94:73–83. https://doi.org/10.1017/S0016672312000274.

Wiggans, G. R., T. A. Cooper, P. M. Vanraden, and J. B. Cole. 2011. Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. J. Dairy Sci. 94:6188–6193. https://doi.org/10.3168/jds.2011-4481.

Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and GWAS. Front. Genet. 7:151. https://doi.org/10.3389/fgene.2016.00151.