Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus

D.A.L Lourenco, Shogo Tsuruta, B.O. Fragomeni, Yutaka Masuda, Ignacio Aguilar, Andres Legarra, J.K. Bertrand, T.S. Amen, L. Wang, D.W. Moser, et al.

Journal of Animal Science
Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

Running head: Single-step genomic BLUP for Angus

Genetic evaluation using single-step genomic BLUP in American Angus[1]

D. A. L. Lourenco*[2], 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*

*Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602

[†]Instituto Nacional de Investigacion Agropecuaria, Canelones, Uruguay 90200

[‡]Institut National de la Recherche Agronomique, UMR1388 GenPhySE, Castanet Tolosan, France 31326

[§]Angus Genetics Inc., St. Joseph, MO 64506

---

[1]This study was partially funded by the American Angus Association (St. Joseph, MO), Zoetis (Kalamazoo, MI), and 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. We gratefully acknowledge the very helpful comments by the two anonymous reviewers.

[2]Corresponding author: danilino@uga.edu

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

ABSTRACT: Predictive ability of genomic EBV when using single-step genomic BLUP (ssGBLUP) in Angus cattle was investigated. Over 6 million records were available on birth weight (**BW**) and weaning weight (**WW**), almost 3.4 million on post-weaning gain (**PWG**), and over 1.3 million on calving ease (**CE**). Genomic information was available on at most 51,883 animals, which included high and low EBV accuracy animals. Traditional EBV was computed by BLUP and genomic EBV by ssGBLUP and indirect prediction based on SNP effects derived from ssGBLUP; SNP effects were calculated based on the following reference populations: ref_2k (high EBV accuracy sires and cows), ref_8k (ref_2k, plus all genotyped ancestors of validation animals), and ref_33k (ref_8k, plus all remaining genotyped animals not in the validation). Indirect prediction was obtained as direct genomic value (DGV) or as an index of DGV and parent average (PA). Additionally, runs with ssGBLUP used the inverse of the genomic relationship matrix calculated by an algorithm for proven and young animals (APY) that uses recursions on a small subset of reference animals. An extra reference subset included 3872 genotyped parents of genotyped animals (ref_4k). Cross-validation was used to assess predictive ability on a validation population of 18,721 animals born in 2013. Computations for growth traits used multiple-trait linear model, and for CE, a bivariate CE-BW threshold-linear model. With BLUP, predictivities were 0.29, 0.34, 0.23, and 0.12 for BW, WW, PWG, and CE, respectively. With ssGBLUP and ref_2k (ref_33k), predictivities were 0.34, 0.35, 0.27, and 0.13 (0.39, 0.38, 0.29, and 0.13), respectively. Low predictivity for CE was due to low incidence rate of difficult calving. Indirect predictions with ref_33k were as accurate as with full ssGBLUP. Using APY and recursions on ref_4k (ref_8k) gave 88% (97%) gains of full ssGBLUP. Genomic evaluation in beef cattle with ssGBLUP is feasible while keeping the models (maternal, multiple trait, threshold) already used in regular BLUP. Gains in predictivity are dependent on

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

the composition of the reference population. Indirect predictions via SNP effects derived from ssGBLUP allow for accurate genomic predictions on young animals, with no advantage of including PA in the index if the reference population is large. With APY conditioning on about 10,000 reference animals, ssGBLUP is potentially applicable to large number of genotyped animals without compromising predictive ability.

Key words: beef cattle, genomic recursion, genomic selection, indirect prediction

INTRODUCTION

Genomic selection in beef cattle has currently been performed with multistep methods, which uses deregressed EBV to estimate SNP effects and then direct genomic value (DGV) for selection candidates based on their genotypes (Meuwissen et al., 2001; Garrick et al., 2009). The main advantage of this approach is that the traditional BLUP evaluation is kept unchanged and genomic selection can be carried out by a separate entity owning genotypes but not phenotypes. Also new animals are easily evaluated if DGV is computed as a sum of marker effects, but not if selection indexes including DGV and parent average (PA) are used.

When both phenotypes and genotypes are available jointly, single-step genomic BLUP (ssGBLUP) (Aguilar et al., 2010) is a simple alternative. This method does not rely on deregressed proofs, properly weighs information from genotyped sires and cows, thus avoiding double-counting of contributions due to relationships and records, and accounts for pre-selection bias of genomically selected parents without phenotypes (Legarra et al., 2014). In ssGBLUP it is also possible to quickly evaluate young genotyped animals without running a complete evaluation that requires several hours to converge. Quick predictions can be calculated indirectly,

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

where genomic predictions for young animals are obtained from SNP effects. It was shown by Wang et al. (2012) that SNP effects can be derived from GEBV solutions from the main ssGBLUP evaluation.

In its current implementation, ssGBLUP uses direct inversion of genomic matrices (Aguilar et al., 2011), which has a cubic cost and a limit of 150,000 animals (Aguilar et al., 2013). Several methods were proposed to overcome that limit (Legarra and Ducrocq, 2012; Fernando et al., 2014; Liu et al., 2014), but none was successful. Recently Misztal et al. (2014) presented a method which uses an approximate inversion of genomic relationships based on recursions on a fraction of the total population; which can be suitable and inexpensive. The first goal of this study was to evaluate the feasibility of ssGBLUP for genomic evaluation in Angus cattle with reference populations of different composition. An additional goal was to evaluate the ability to predictive GEBV with genomic recursions and with indirect prediction for young animals.

MATERIAL AND METHODS

Datasets from American Angus Association (AAA) were available for this study that included growth traits and calving ease (CE). Growth traits included birth weight (BW), weaning weight (WW), and post-weaning gain (PWG). As the data were obtained from existing databases, Animal Care and Use Committee approval was not obtained for this study.

Data

Over 6 million phenotypes were available for BW and WW, almost 3.4 million for PWG, and over 1.3 million for CE. Whereas BW, WW, and PWG are continuous traits, CE is a categorical trait with 5 calving scores, where 5 is abnormal delivery and is excluded. Because few animals had scores 3 and 4, these scores were combined into category 2, which resulted in 93% of

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

animals with score 1 and 7% with score 2. The number of animals in the pedigree for evaluation of growth traits was 8,236,425, and for CE was 8,025,676.

For evaluation of growth traits, 81,878 animals were genotyped for 54,609 SNP from the BovineSNP50k v2 BeadChip (Illumina Inc., San Diego, CA). Currently, no genotyping strategy is applied by AAA; therefore, the members can choose which animals are being genotyped, and most of them are young. A total of 29,995 genotyped animals were young without phenotypes for any of the 3 traits, which caused them to have genotypes excluded from this study. If the number of genotyped animals is relatively large, young genotyped animals without phenotypes in the dataset give very small contribution to their relatives' evaluation (Misztal et al., 2014). After removing SNP with unknown position or located on sex chromosomes and running a general quality control analysis, genotypes on 38,528 SNP markers were available for 32,465 males and 19,418 females born from 1977 to 2013; therefore, the maximum number of genotyped animals used in all analyses on growth traits was 51,883. For CE evaluation, a genotyping set with 72,069 animals was available, but only genotypes on 40,546 animals born from 1977 to 2013 (26,074 males and 14,472 females) were used for the same reason above. The number of SNP that passed the general quality control for this dataset was 38,568.

For this study, the animals were then split into training and validation populations according to year of birth. Thus, all 18,721 (13,166) genotyped animals born in 2013 were chosen to be in the validation population for growth (CE) traits and had their phenotypes removed from the evaluations. The pedigree relationship between training and validation populations ranged from 0 to 0.82, with an average relationship of 0.09.

Model

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

Traditional and genomic evaluations were performed for growth traits and CE. A multivariate linear animal model was used for growth traits as:

$$\mathbf{y}_t = \mathbf{Xb} + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\mathbf{m} + \mathbf{Z}_3\mathbf{p} + \mathbf{e} \qquad [1]$$

where t is for each one of BW, WW, PWG; **y**, **b**, **u**, **m**, **p**, and **e** are vectors of phenotypes, fixed effect of contemporary group, additive direct genetic effect, additive maternal genetic effect, maternal permanent environmental effect, and random residuals, respectively; **X**, $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ are incidence matrices for **b**, **u**, **m**, and **p**, respectively. All random effects were present for WW, but only **u**, **m**, and **e** for BW, and **u** and **e** for PWG.

A bivariate threshold-linear animal model was used to model CE jointly with BW:

$$\mathbf{Y}_c = \mathbf{Xb} + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\mathbf{m} + \mathbf{e} \qquad [2]$$

where c is for BW and CE; **y**, **b**, **u**, **m** and **e** are vectors of phenotypes, fixed effects of contemporary group, sex, age of dam (only for CE), and sex by age of dam interaction (only for CE), additive direct genetic effect, additive maternal genetic effect, and random residuals, respectively; **X**, $\mathbf{Z}_1$, and $\mathbf{Z}_2$ are incidence matrices for **b**, **u**, and **m**, respectively. According to Ramirez-Valverde et al. (2001) when BW is available, bivariate threshold-linear models including CE and BW are a better alternative than a single-trait threshold model to evaluate CE, especially if the population has animals with different levels of EBV accuracy. From this model, only results for CE are discussed, whereas results for BW are from the multiple trait linear model for growth traits. Heritabilities for all traits were calculated by AAA using the same models as in [1] for BW, WW, and PWG; and in [2] for CE. For our study, the values were then provided by AAA and ranged from 0.12 to 0.41 (Table 1).

Analyses

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

Three different genomic analyses were performed using ssGBLUP (Aguilar et al., 2010; Christensen and Lund, 2010) as implemented in BLUP90IOD program (http://nce.ads.uga.edu/wiki/BLUPmanual). Compared to BLUP, in ssGBLUP the inverse of the numerator relationship matrix $\mathbf{A}^{-1}$ is replaced by matrix $\mathbf{H}^{-1}$ defined as follows:

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

where **G** is the genomic relationship matrix. The computations used default options in BLUP90IOD. In all analyses the validation population was defined as genotyped animals born in 2013 with phenotypes excluded.

First analysis: ssGBLUP with different reference populations. Different reference populations were defined according to EBV accuracy calculated with the ACCF90 program (http://nce.ads.uga.edu/wiki/BLUPmanual), which uses the concept of prediction error variance and reflects the standard error of EBV for each individual. The objective was to investigate the influence of different groups of reference animals on genomic predictions, and possibly to guide genotyping strategy. The current trend in livestock genomics is to genotype young animals; however, more important animals give more information to the evaluations. For growth traits (CE), the first reference population was composed of 1,628 (1,541) top bulls with EBV accuracy for BW $\geq$ 0.85; which we will refer hereinafter as "ref_bulls". As BW was present in models for growth and CE evaluations, using its EBV accuracy for selecting top bulls helped to compose sets with proportional number of animals. In this case, the **G** matrix was composed of animals in the reference population and also animals in the validation population; the last had 18,721 animals for growth traits and 13,166 for CE. The second reference population was composed of the top bulls and also top cows that had an EBV accuracy for BW $\geq$ 0.85; which we will refer as

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

ref_2k. The number of top cows was small and only 268 were added for the growth trait analysis and 323 for CE. The third reference population was composed of top bulls, top cows, and all other genotyped animals born from 1977 to 2012 (we will refer as ref_33k). This group had a total of 33,162 animals for growth and 27,380 for CE, with an average EBV accuracy for BW of 0.77 (± 0.16). For the latter analysis, the **G** matrix was composed of the maximum number of 51,883 genotyped animals for growth analysis and 40,546 for analysis of CE.

Second analysis: ssGBLUP with indirect predictions for young animals. With the increasing number of genotyped heifers and steers in dairy and beef, the genomic methods should be able to provide predictions for young animals without phenotypes in a quick run, externally to the official evaluations. This concept is introduced here as indirect ssGBLUP, and basically mimics the mixed model equations. It would be advantageous from different perspectives: to evaluate young animals mainly for traits that are measured later in life, after the selection decisions are made; and to reduce computing costs because the dimension of **G** would not increase in the same proportion as the number of genotyped animals.

In order to explain how it works, consider the equation for the GEBV of a single individual in ssGBLUP (VanRaden and Wiggans, 1991; Aguilar et al., 2010; Lourenco et al., 2015):

$$GEBV = w_1PA + w_2YD + w_3PC + w_4DGV - w_5PP$$

where PA is Parent Average, YD is Yield Deviation (phenotypes adjusted for model effects other than additive genetic and error), PC is Progeny Contribution, DGV is direct genomic value (corresponding to $\mathbf{G}^{-1}$), PP is the pedigree prediction based on the subset of genotyped animals from **A** (corresponding to $\mathbf{A}_{22}^{-1}$) and w_1 to w_5 are weights that sum to 1. In the case of young animals with no progeny or own performance record, YD=PC=0 and w_2=w_3=0. In this case, for

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

155 individual i, $PA_i=(GEBV_s+GEBV_d)/2$; $DGV_i= -\frac{\sum_{j,j\neq i} g^{ij}GEBV^j}{g^{ii}}$; $PP=-\frac{\sum_{j,j\neq i} a_{22}^{ij}GEBV^j}{a_{22}^{ii}}$ and $w_1=\frac{2}{den}$,

156 $w_4=\frac{g^{ii}}{den}$, $w_5=\frac{a_{22}^{ii}}{den}$, where *den* is the denominator that equals to $2+\left(g^{ii}-a_{22}^{ii}\right)$; g^{ij} (a_{22}^{ij}) is an

157 element of $\mathbf{G}^{-1}$ ($\mathbf{A}_{22}^{-1}$) corresponding to relationships between animal *i* and *j*; *s* and *d* correspond to

158 sire and dam, respectively. If all individuals are genotyped, then PA=PP and GEBV reduces to

159 DGV.

160 For ssGBLUP with indirect predictions, SNP effects can be calculated using the current run

161 of ssGBLUP with all but young animals, and genomic predictions for young animals are

162 obtained by multiplying the SNP content by SNP effect to obtain DGV; a more complete GEBV

163 can also be available through a selection index that combines DGV and PA. The flow for indirect

164 predictions in ssGBLUP is:

165 1) Run ssGBLUP with a reference population to obtain GEBV. In this step, 3 reference

166 populations were tested:

167 a) ref_2k: reference population with top bulls and top cows (n=1,896);

168 b) ref_8k: reference population with all parents that were genotyped (n=8,285), this

169 includes ref_2k;

170 c) ref_33k: reference population with all genotyped animals born up to 2012 (n=33,162),

171 this includes ref_8k;

172 2) Split GEBV into all the components shown before, where DGV for an animal i in the

173 reference population is calculated as below (Aguilar et al. (2010):

$$DGV_i=-\frac{\sum_{j,j\neq i} g^{ij}GEBV^j}{g^{ii}}$$

174 with all elements previously defined.

175 3) Calculate SNP effects using DGV from the reference population:

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

$$\hat{\mathbf{u}} = \mathbf{DZ}'\mathbf{G}^{-1}(\mathbf{DGV})$$

where $\hat{\mathbf{u}}$ is a vector of estimated SNP effects, **D** is a diagonal matrix of weights (standardized variances) for SNP (identity matrix in this case), and **Z** is a matrix of centered genotypes for each animal (VanRaden, 2008). A similar approach that uses GEBV instead of DGV to calculate SNP effects was proposed by Wang et al. (2012). However, for numerical purposes this involves approximations as **G** matrix is formed as $\mathbf{G}=0.95\mathbf{ZDZ}'+0.05\mathbf{A}_{22}$ (Aguilar et al., 2010). This is done as a default approach to avoid singularity problems and may result in negligible error as shown later.

4) Calculate DGV for young genotyped animals ($\mathbf{DGV}_y$):

$$\mathbf{DGV}_y = \mathbf{Z}_y\hat{\mathbf{u}}$$

where $\mathbf{DGV}_y$ and $\mathbf{Z}_y$ are direct genomic values and a matrix of centered genotypes for young animals not included in ssGBLUP evaluation, respectively.

5) Combine DGV_y with PA for young genotyped animals:

$$GEBV_y \approx w_1 PA + w_4 DGV_y$$

where $GEBV_y$ is GEBV obtained via indirect predictions for young animals, w_1 and w_4 are weights identical for all animals and calculated based on covariances between DGV_y and PA as:

$$\begin{bmatrix} w_1 \\ w_4 \end{bmatrix} = \begin{bmatrix} \sigma^2_{PA} & \sigma_{PA,DGV_y} \\ \sigma_{DGV_y,PA} & \sigma^2_{DGV_y} \end{bmatrix} \begin{bmatrix} \sigma^2_{PA} \\ \sigma^2_{DGV_y} \end{bmatrix}$$

Note this is an approximation which ignores PP. In general, PP includes part of PA explained by DGV. When all animals are genotyped, PP and PA cancel out, with approximate cancellation when parents of an animal are genotyped. When an animal is unrelated to a genotyped population, PP=0. Fixed weights in the index account for an

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

195 average relationship of all young animals to a genotyped population. It is possible to
196 create different indices based on the number of genotyped parents (VanRaden et al.,
197 2012).

198 The ssGBLUP with indirect prediction allows calculation of DGV or GEBV for young
199 genotyped animals, with lower computing cost compared to a full ssGBLUP where young
200 animals are explicitly included.

201 ***Third analysis: ssGBLUP with G inverted by a recursive algorithm.*** When the number of
202 genotyped animals is large and there is a desire for using all of them in ssGBLUP evaluations to
203 get direct predictions for all, including young animals, an algorithm that splits genotypes into
204 proven and young animals and uses recursion to approximate the inverse of the **G** matrix was
205 proposed by Misztal et al. (2014). This algorithm is known as APY, and $\mathbf{G}^{-1}$ containing all
206 genotyped animals can be expressed as:

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}_{pp}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{pp}^{-1}\mathbf{G}_{py} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_g^{-1} \begin{bmatrix} -\mathbf{G}_{yp}\mathbf{G}_{pp}^{-1} & \mathbf{I} \end{bmatrix}$$

207 where the subscript pp stands for proven animals and py for the covariance between proven and
208 young animals; each element of $\mathbf{M_g}$ is obtained (for the i^{th} young animal) as
209 $m_{g,i} = g_{ii} - \mathbf{G}_{ip}\mathbf{G}_{pp}^{-1}\mathbf{G}_{pi}$ and is called genomic Mendelian sampling. In APY, the only direct
210 inversion needed is for part of **G** that contains relationships among proven animals ($\mathbf{G}_{pp}$),
211 whereas all other coefficients are obtained through recursions.

212 For this analysis, four definitions of proven animals were tested that included the 3
213 definitions used for indirect predictions (ref_2k, ref_8k, and ref_33k), plus one more definition
214 where 3,872 genotyped parents of genotyped animals were considered as proven (ref_4k). This

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

last group was added to test if proven animals would have strong links with the young genotyped population.

The greatest advantages of this algorithm are the reduction of computing cost, which is still cubic for proven animals, but can be linear for young animals; and the possibility of using large amounts of genotyped animals in ssGBLUP evaluations. The secondary advantage is numerical stability as the regular **G** matrix is singular when the number of animals is greater than the number of SNP markers and cannot be inverted without blending with $\mathbf{A}_{22}$.

Validation

The ability to predict future phenotypes was the validation method chosen for this study. This method is based on Legarra et al. (2008), and predictive ability for traditional and genomic evaluations for animals born in 2013 was calculated as the correlation between (G)EBV and phenotypes corrected for fixed effects (y-Xb):

$$r = cor[(G)EBV, y - Xb]$$

The predictive ability or predictivity is used as an approach to compare the methods applied in this paper. For all analyses, the validation groups were kept the same to make comparisons easier. Validations involved 18,721 animals for growth traits and 16,133 animals for CE. Predictivity calculated with EBV in the above formula was the benchmark used to compare the gain in predictive ability due to genomics, and predictivity calculated with GEBV was used to compare the genomic methods previously described. Prediction accuracy could be described as r/h, where h is square root of heritability; however, prediction accuracy can be overestimated if heritabilities are obtained by simplified models as the ones used by AAA.

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

RESULTS AND DISCUSSION

ssGBLUP with different reference populations

Predictive ability on young animal when using several reference populations is shown in Table 2. Using only top bulls as a reference population (ref_bulls) increased predictivity relative to BLUP by 0.05 for BW, 0.01 for WW, 0.04 for PWG, and 0.01 for CE. Addition of top cows to the reference population (ref_2k) did not increase the predictivity for any trait. This could be due to the small number of animals added and also because daughters of those cows already contributed through the inclusion of bulls. Addition of around 31,000 animals to the reference population provided an additional increase in predictivity of 0.05 for BW, of 0.03 for WW and of 0.02 for PWG. However, no additional increase was observed for CE by adding extra 27,000 genotyped animals, of which about 7,000 had phenotypes for that trait.

The addition of 31,000 animals with few or no progeny led to the same increase of predictivity as using only the top bulls for BW, led to an increase of 3 times for WW and an increase of 0.5 times for PWG. Among the 31,000 extra animals, almost all had phenotypes for BW and WW, but only 24,000 had phenotypes for PWG. Evidently, the composition of reference population is also a factor that influences predictivity of GEBV besides the reference population size. Thus, genotyping strategy should take into account genotyping more important and maybe older animals with more information (higher EBV accuracy) along with genotyping large amounts of young animals.

Previous studies showed that prediction accuracies or predictive ability are biased downward by selection (Bijma, 2012; Lourenco et al., 2015). In our study, it appears that selection for proven bulls was much stronger for WW than for PWG (lower increase in predictivity with twice the phenotypic data at similar heritability) but there was a small selection on genotyped animals

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

with own records (approximately twice the increase of predictivity with twice the phenotypic data). It may be hard to calculate the amount of bias in livestock species, including beef cattle, as the selection process is sequential and affected by both genetic correlations and specific indexes used for selection.

Low predictivity for CE in this study is due to lower heritability combined with limited recording for this trait and a low incidence of difficult calving. Additionally, very few genotyped animals had a difficult calving, perhaps because animals from a difficult calving are unlikely to be retained for breeding and therefore would not be genotyped on a regular basis. Higher predictivity and impact of genomic selection for CE could be expected in breeds with higher incidence of calving problems.

Because the increase in predictivity for CE was very small compared to predictivity of traditional evaluations, indirect predictions and APY were not tested for this trait.

In this paper, only predictivity for the direct genetic effect is shown; however, models for BW and WW included maternal effect, which is also important in genetic evaluations. We attempted to derive formulas for predictivity of maternal effects, unsuccessfully. Such predictivity can be hard to assess because the maternal effect occurs one generation back, which means that the corrected phenotype of animal i should be correlated with the maternal effect of the dam of animal i. But, dams usually have more than one progeny and there is genetic correlation between direct and maternal for BW, which makes derivations difficult. Lourenco et al. (2013) used simulated data that mimicked a beef cattle population and showed that the gain for the maternal effect with ssGBLUP is as high as for the direct effect.

ssGBLUP with indirect predictions for young animals

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

Predictive ability for indirect prediction via conversion of DGV into SNP effects is shown in Figure 1. When the reference population included top bulls and top cows (ref_2k), the predictivity of indirect DGV_y was lower than predictivity for traditional EBV for the three traits (0.23 vs. 0.29 for BW; 0.28 vs. 0.34 for WW; 0.19 vs. 0.23 for PWG). Predictivity for $GEBV_y$ calculated as an index of indirect DGV_y with PA was higher than those for EBV for the three traits (0.31 vs. 0.29 for BW; 0.36 vs. 0.34 for WW; 0.24 vs. 0.23 for PWG), however, this predictivity was lower than the ones from full ssGBLUP (except for WW). With larger reference population (ref_8k), all indirect DGV_y were similar or more accurate than EBV, and the index had similar predictivity as the full ssGBLUP. With the largest reference population (ref_33k), all indirect DGV_y were almost as accurate as GEBV from full ssGBLUP, with the index marginally improving predictivity for WW. This marginal improvement for WW may be caused by the use of less than optimal genetic parameters, e.g., zero covariance between direct and maternal effects (to reduce computing costs). The DGV_y obtained with ref_33k reference population were more accurate than GEBV from full ssGBLUP obtained with ref_8k reference population.

Although predictivity of indirect predictions when using ref_33k was similar to predictivity from full ssGBLUP, it does not mean that predictions have the same average. The reason for that is the different sources of information used to calculate indirect predictions. Correlations between GEBV and indirect predictions are a good tool to assure that the latter can be used for interim evaluations. Correlations between GEBV from full ssGBLUP and DGV_y or $GEBV_y$ from indirect predictions are shown in Table 3. On average, correlations with DGV_y were 0.73, 0.89, and 0.96 for ref_2k, ref_8k, and ref_33k, respectively. Higher correlations were observed between GEBV and $GEBV_y$, with values for the three reference sets being 0.89, 0.95, and 0.97, respectively. Those results endorse the use of a reference population of size close to 33,000

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

303 animals for this American Angus dataset. By doing that, indirect predictions are as accurate as
304 predictions including genotypes for young animals in the evaluation (full ssGBLUP).
305 For young animals, GEBV = w_1PA + w_4DGV – w_5PP, with all weights summing to 1.0
306 (VanRaden and Wiggans, 1991; VanRaden et al., 2009; Aguilar et al., 2010). When the number
307 of genotyped animals is small, w_4 is small and ignoring PA reduces predictivity. Using an index
308 with PA improves the predictivity, however, PP is ignored and computed weights w_1 and w_4 are
309 approximate. When the number of genotyped animals is large, w_4 is close to 1.0, and ignoring
310 PA marginally reduces the predictivity for some traits. Therefore, the indirect prediction via
311 DGV is accurate when SNP effects are derived from ssGBLUP with sufficient size of the
312 reference population.
313 Neglecting PP seems to have no considerable effect in this population, because predictivity
314 of indirect predictions was very similar to predictivity from full ssGBLUP. Neglection of PP
315 indirectly means adjusting PA for an average PP. VanRaden et al. (2012) used different weights
316 for animals based on the number of genotyped parents, which better accounts for PP.
317 A study by Wiggans et al. (2014) used SNP effects from previous monthly genomic multi-
318 step evaluations to calculate preliminary GEBV for young genotyped animals. The objective was
319 to have daily or weekly genomic evaluations for US dairy cattle and reduce the time between
320 having DNA samples and predictions from a monthly official evaluation. Their reference set
321 contained all genotyped animals with phenotypes (about 597,000; corresponding to ref_33k in
322 our study) and correlations between preliminary and official evaluations were higher than 0.99
323 for Holsteins, but smaller for other breeds with a smaller number of genotyped animals. Further
324 research with different species will be critical in determining the sufficient size of the reference
325 population for indirect predictions in order to achieve high predictivity. It may be related to

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

effective population size, number of independent SNP (Pintus et al., 2013), and relationships between reference and validation populations as in multi-step methods. Although indirect predictions via ssGBLUP use a concept similar to multi-step methods for young genotyped animals, indirect predictions via ssGBLUP may be more accurate than multistep predictions because the latter are affected by approximations involved in deregressions and possible double-counting of phenotypic information.

For young animals, indirect predictions via SNP effects from ssGBLUP seems a viable alternative as it can be done separately from the full evaluation. As SNP effects are calculated based on trait GEBV or DGV, indirect predictions are easily obtained for multi-trait models, as done in this study; multi-breed and crossbred evaluations are possible when the G matrix is able to account for information on all breeds. However, if young animals and particularly full-sibs are intensively selected, selection on the Mendelian sampling will not be accounted for, leading to pre-selection bias (Patry and Ducrocq, 2011). Analyses by ssGBLUP with all genotypes subject to selection are expected to account for pre-selection (VanRaden and Wright, 2013), because selection is accounted for when all information used for selection is included in the model (Henderson, 1975).

Comments on SNP weighting and SNP selection

The way SNP effects are calculated in ssGBLUP allows for inclusion of different weights for SNP: $\hat{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{G}^{-1}$(DGV), with weights for **G** fit into the diagonal matrix **D**. Those weights can be calculated through an iterative process, or external weights can be used as input for ssGBLUP (Wang et al., 2012; Su et al., 2014). Weighting **G** seems to be a reasonable approach to achieve higher prediction accuracy, especially in the presence of "major" SNP. Sun et al. (2011) showed

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

higher prediction accuracy when using weighted **G** in regular GBLUP compared to BayesB. For some traits, SNP weighting or SNP selection in ssGBLUP also gave additional prediction accuracy (Wang et al., 2014). In fact, when weights are different per trait, this precludes the use of multiple traits unless the model includes one common additive effect and specific additive effects for individual traits. In practice and especially under a selection index, gains from a multiple trait analysis can overcome losses due to not fitting "major" SNP. Also, when the number of genotyped animals increases, the rate of gain in reliability increases at a slower pace (VanRaden et al., 2011); therefore, weighting SNP may no longer have a big impact on prediction accuracy (Winkelman et al., 2015).

ssGBLUP with G inverted by a recursive algorithm

Predictive ability of GEBV when the inverse of **G** is computed with APY is shown in Figure 2. When the recursions were conditioned on ref_2k, ref_4k, ref_8k, and ref_33k, the procedure accounted for 67%, 88%, 97%, and 100% of predictivity gains of ssGBLUP over BLUP, respectively. Therefore, in ssGBLUP, using genomic recursion to invert **G** while conditioning on enough number of animals, in this case about 8,000, has the same prediction power as **G** using direct inversion. The amount of memory necessary for APY $\mathbf{G}^{-1}$ using ref_2k, ref_4k, ref_8k, and ref_33k was approximately 0.8, 1.6, 3.2, and 13.7 Gbytes, respectively, whereas the amount of memory for the regular $\mathbf{G}^{-1}$ is 21.6 Gbytes. Therefore, using APY $\mathbf{G}^{-1}$ makes computations less costly and faster.

Tests involving 100,000 genotyped Holsteins with recursions conditioned on more than 15,000 animals resulted in practically identical GEBV compared to the regular inversion but with a better convergence rate (Fragomeni et al., 2015) indicating that APY has good predictive and

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

numerical properties. They suggested that the necessary number of animals being conditioned is proportional to the number of independent chromosome segments, which is a function of an effective population size.

The main advantages of APY are low computing costs and numerical stability. With conditioning on 8,000 animals, for example, the only inverse required is for a block of **G** for 8,000 animals, and additional genotypes require only linear storage and computations. Subsequently, computations with a large number of genotyped animals may be feasible with similar predictivity as in the regular inversion. APY would be the algorithm of choice for regular evaluations with very large number of genotyped animals.

CONCLUSIONS

Genomic evaluation in beef cattle using single-step genomic BLUP is feasible for either linear or categorical traits. Gains in predictive ability over BLUP are dependent on the size and composition of the reference population, and are large for growth traits and small for CE. With a sufficient number of animals in the reference population, indirect prediction for young animals via SNP effects provides similar predictivity to full single-step genomic BLUP, allowing for quick genomic predictions without running a complete evaluation. Use of the algorithm for proven and young animals in single-step genomic BLUP allows for incorporation of large number of genotyped animals at low cost without compromising the predictive ability.

LITERATURE CITED

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

389 Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic:
390 A unified approach to utilize phenotypic, full pedigree, and genomic information for
391 genetic evaluation of Holstein final score. J. Dairy Sci. 93: 743–752.
392 Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic
393 relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed.
394 Genet. 128: 422–428.
395 Aguilar, I., A. Legarra, S. Tsuruta, and I. Misztal. 2013. Genetic evaluation using unsymmetric
396 single step genomic methodology with large number of genotypes Interbull Bull. 47:
397 222–225.
398 Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do
399 not reflect the correlation between true and estimated breeding values in selected
400 populations. J. Anim. Breed. Genet. 129: 345–358.
401 Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not
402 genotyped. Genet. Sel. Evol. 42: 2.
403 Fernando, R. L., J. C. M. Dekkers, and D. Garrick. 2014. A class of Bayesian methods to
404 combine large numbers of genotyped and non-genotyped animals for whole-genome
405 analyses. Genet. Sel. Evol. 46: 50.
406 Fragomeni, B. O., D. A. L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J.
407 Lawlor, and I. Misztal. 2015. Use of genomic single-step genomic BLUP with a large
408 number of genotypes. J. Dairy Sci. *(accepted)*.
409 Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values
410 and weighting information for genomic regression analyses. Genet. Sel. Evol. 41: 55.

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

411 Henderson, C. R. 1975. Best linear umbiased estimation and prediction under a selection model.
412 Biometrics 31: 423–447.
413 Legarra, A., C. R. Granie, E. Manfredi, and J. M. Elsen. 2008. Performance of genomic selection
414 in mice. Genetics 180: 611–618.
415 Legarra, A., and V. Ducrocq. 2012. Computational strategies for national integration of
416 phenotypes, genomic, and pedigree data in single-step best linear unbiased prediction. J.
417 Dairy Sci. 95: 4629–4645.
418 Legarra, A., O. F. Chistensen, I. Aguilar, and I. Misztal. 2014. Single step, a general approach
419 for genomic selection. Livest. Prod. Sci. 166: 54–65.
420 Liu, Z., M. E. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with
421 direct estimation of marker effect. J. Dairy Sci. 97: 5833–5850.
422 Lourenco, D. A. L., I. Misztal, H. Wang, I. Aguilar, S. Tsuruta, and J. K. Bertrand. 2013.
423 Prediction accuracy for a simulated maternally affected trait of beef cattle using different
424 genomic evaluation models. J. Anim. Sci. 91: 4090–4098.
425 Lourenco, D. A. L., I. Misztal, B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J.
426 Hawken, and A. Legarra. 2015. Accuracies of males and females with genomic
427 information on males, females, or both: a broiler chicken example. Genet. Sel. Evol.
428 *(Submitted)*.
429 Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value
430 using genome-wide dense marker maps. Genetics 157: 1819–1829.
431 Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the
432 genomic relationship matrix. J. Dairy Sci. 97: 3943–3952.

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

433 Patry, C., and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic
434 preselection in dairy cattle. J. Dairy Sci. 94: 1011–1020.
435 Pintus, M. A., E. L. Nicolazzi, J. B. C. H. M. V. Kaam, S. Biffani, A. Stella, G. Gaspa, C.
436 Dimauro, and N. P. P. Macciotta. 2013. Use of different statistical models to predict
437 direct genomic values for productive and functional traits in Italian Holsteins J. Anim.
438 Breed. Genet. 130: 32–40.
439 Ramirez-Valverde, R., I. Misztal, and J. K. Bertrand. 2001. Comparison of threshold vs linear
440 and animal vs sire models for predicting direct and maternal effects on calving difficulty
441 in beef cattle. J. Anim. Sci. 79: 333–338.
442 Su, G., O. F. Christensen, L. Janss, and M. S. Lund. 2014. Comparison of genomic predictions
443 using genomic relationship matrices built with different weighting factors to account for
444 locus-specific variances. J. Dairy Sci. 97: 6547–6559.
445 Sun, X., R. L. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2011. An iterative approach for
446 efficient calculation of breeding values and genome-wide association analysis using
447 weighted genomic BLUP. J. Anim. Sci. 89 (E-Suppl.2): 28. (Abstr.).
448 VanRaden, P. M., and G. R. Wiggans. 1991. Deviation, calculation, and use of national animal
449 model information. J. Dairy Sci. 74: 2737–2746.
450 VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:
451 4414–4423.
452 VanRaden, P. M., C. P. VanTassel, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F.
453 Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for
454 North American Holstein bulls. J. Dairy Sci. 92: 16–24.

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

455 VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluations
456 with many more genotypes. Genet. Sel. Evol. 43: 10.
457 VanRaden, P. M., J. R. Wright, and T. A. Cooper. 2012. Adjustment of selection index
458 coefficients and polygenic variance to improve regressions and reliability of genomic
459 evaluations. J. Dairy Sci. 95 (Suppl. 2): 520 (Abstr.).
460 VanRaden, P. M., and J. R. Wright. 2013. Measuring genomic pre-selection in theory and in
461 practice. Interbull Bull. 47: 147–150.
462 Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association
463 mapping including phenotypes from relatives without genotypes. Genet. Res. 94: 73–83.
464 Wang, H., I. Misztal, I. Aguilar, A. Legarra, R. L. Fernando, Z. Vitezica, R. Okimoto, T. Wing,
465 R. Hawken, and W. M. Muir. 2014. Genome-wide association mapping including
466 phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body
467 weight in broiler chickens. Front. Genet. 5: 134.
468 Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2014. Technical note: Rapid calculation of
469 genomic evaluations for new animals. J. Dairy Sci. 98: 2039–2042.
470 Winkelman, A. M., D. L. Johnson, and B. L. Harris. 2015. Application of genomic evaluation to
471 dairy cattle in New Zealand. J. Dairy Sci. (*In press*).
472

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

473 **Table1.** Heritability (h^2) and general statistics for growth traits and CE

Trait[1]	h^2	Number of records	Average (kg)	SD (kg)	Number of genotyped animals with records
BW	0.41	6,189,661	36.47	4.45	50,784
WW	0.20	6,890,625	263.13	44.63	51,830
PWG	0.20	3,387,252	162.25	67.00	36,196
CE	0.12	1,310,684	-	-	10,558
easy	-	1,215,571	-	-	10,228
difficult	-	95,113	-	-	330

474 [1] BW = birth weight; WW = weaning weight; PWG = post-weaning gain; CE = calving ease.

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

475 **Table 2**. Predictive ability of future phenotypes for young genotyped animals born in 2013

Trait[1]	Animals in validation	BLUP	ssGBLUP[2]		
			ref_bulls[3]	ref_2k	ref_33k
BW	18,721	0.29	0.34	0.34	0.39
WW	18,721	0.34	0.35	0.35	0.38
PWG	18,721	0.23	0.27	0.27	0.29
CE	13,166	0.12	0.13	0.13	0.13

476 [1] BW = birth weight; WW = weaning weight; PWG = post-weaning gain; CE = calving ease.
477 [2] Single-step genomic BLUP (ssGBLUP) included genotypes for reference and validation
478 populations, but phenotypes for validation animals were removed. Predictive ability was
479 calculated as correlation between corrected phenotypes and genomic EBV.
480 [3] ref_bulls is a reference populations that contains top bulls, ref_2k contains top bulls and top
481 cows, and ref_33k contains all genotyped animals born up to 2012.

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

Table 3. Correlations between GEBV from full ssGBLUP and DGV_y or $GEBV_y$ from indirect predictions.

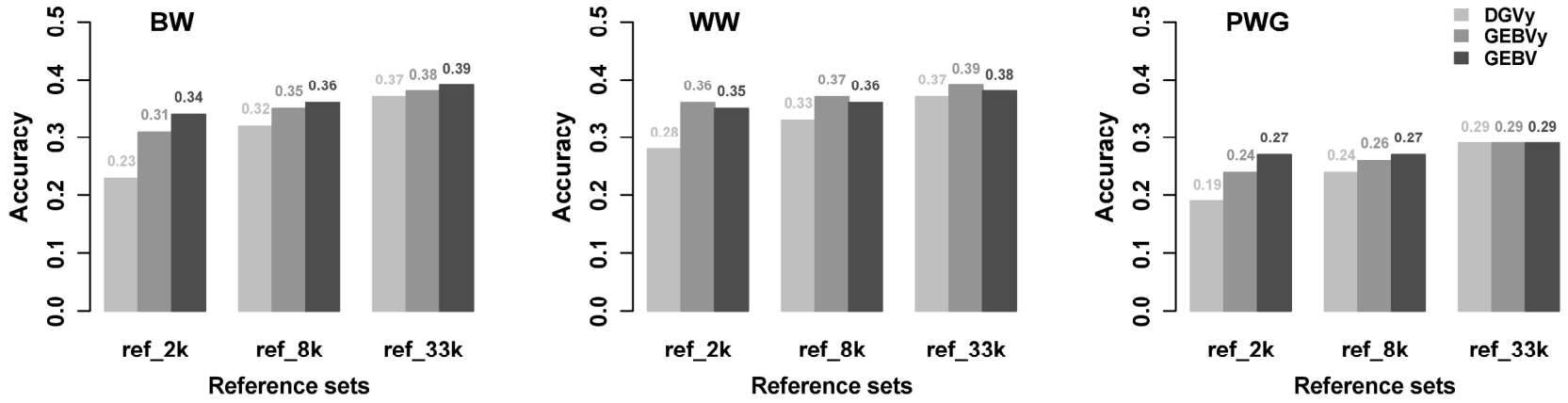
Trait	Indirect Prediction[1]	ref_2k[2]	ref_8k	ref_33k
BW	DGV_y	0.66	0.87	0.96
	$GEBV_y$	0.85	0.94	0.97
WW	DGV_y	0.75	0.89	0.95
	$GEBV_y$	0.90	0.95	0.97
PWG	DGV_y	0.78	0.90	0.96
	$GEBV_y$	0.91	0.96	0.97

[1] DGV_y is direct genomic value; $GEBV_y$ is the indirect genomic EBV obtained by an index combining parent average and DGV_y.

[2] ref_2k is a reference populations that contains top bulls and top cows, ref_8k contains all parents that were genotyped, and ref_33k contains all genotyped animals born up to 2012.

Journal of Animal Science

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836



488 **Figure 1.** Predictive ability of indirect predictions on 18,721 young genotyped animals when using reference populations ref_2k,
489 ref_8k, and ref_33k animals to run single-step genomic BLUP (ssGBLUP) and derivate SNP effects; ref_2k is a reference populations
490 that contains top bulls and top cows, ref_8k contains all parents that were genotyped, and ref_33k contains all genotyped animals born
491 up to 2012. DGV_y is direct genomic value; $GEBV_y$ is the indirect genomic EBV obtained by an index combining parent average and
492 DGV_y; GEBV is genomic predictions obtained directly from ssGBLUP when genotypes on reference and validation animals were
493 considered together in evaluations. Predictive ability was calculated as correlation between corrected phenotypes and genomic EBV.

Journal of Animal Science Accepted paper, posted 04/06/2015. doi:10.2527/jas2014-8836

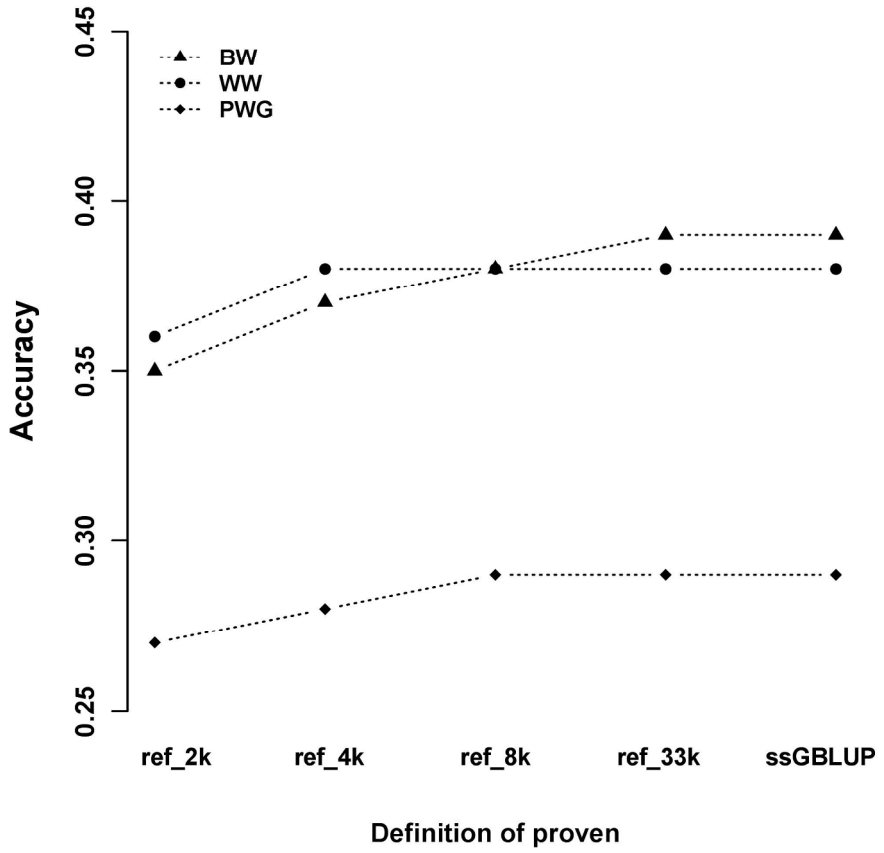


Figure 2. Predictive ability of GEBV for 18,721 young genotyped animals when using APY (algorithm for proven and young animals) to invert **G** matrix (genomic-based relationship matrix) with different definitions of proven animals: ref_2k, ref_4k, ref_8k, and ref_33k; ref_2k is a reference populations that contains top bulls and top cows, ref_4k contains genotyped parents of genotyped animals, ref_8k contains all parents that were genotyped, and ref_33k contains all genotyped animals born up to 2012. Predictive ability was calculated as correlation between corrected phenotypes and genomic EBV. Predictions in single-step genomic BLUP (ssGBLUP) are obtained through direct inversion of **G**.