Single-Step Genomic Evaluations with 570K Genotyped Animals in US Holsteins

***Y. Masuda*[1], *I. Misztal*[1], *S. Tsuruta*[1], *D. A. L. Lourenco*[1], *B. O. Fragomeni*[1], *A. Legarra*[2], *I. Aguilar*[3], *and T. J. Lawlor*[4]**

[1] University of Georgia, Athens, GA 30602
[2] INRA, UR631 SAGA, BP 52627, 31326 Castanet-Tolosan Cedex, France
[3] Instituto Nacional de Investigación Agropecuaria, Canelones, Uruguay 90200
[4] Holstein Association USA Inc., Brattleboro, VT 05301

Abstract

The objectives of this study were to implement and evaluate the “Algorithm for proven and Young” (APY) for inversion of the genomic relationship matrix (**G**) in single-step genomic BLUP (ssGBLUP). Phenotypic data included 11,626,576 final scores on 7,093,380 US Holsteins and genotypes were available for 569,404 animals. Daughter deviations for young genotyped bulls with no classified daughters in 2009 but with at least 30 classified daughters in 2014 were computed using BLUP with all the phenotypes and pedigrees. Genomic predictions (GEBV) were obtained by ssGBLUP using phenotypes up to 2009. We calculated the **G** inverse with APY based on genomic recursions on a subset of “base” animals. We tested several subsets including 9,406 bulls with at least 1 daughter, 9,046 bulls and 1052 dams, 9,046 bulls and 7,422 classified cows, and random samples of 5,000, 10,000, 15,000, 20,000, and 30,000 animals. Validation reliability was calculated as R^2 with a linear regression of daughter deviations on GEBV for young genotyped bulls. The reliabilities were 0.39 with 5,000 randomly chosen base-animals, 0.45 with base-animals including bulls and cows, and 0.44 with the remaining subsets. Setting up the **G** inverse for all the genotypes with 10,000 base-animals took 1.3 hours and 57GB of memory. Genomic predictions with **G** inverse are accurate when the number of base animals is at least 10,000. Single-step genomic BLUP using the **G** inverse via APY is applicable to populations with a large number of genotyped animals.

Key words: APY, ssGBLUP, validation

Introduction

Single-step genomic BLUP (**ssGBLUP**) is a tool for genomic evaluations with several advantages over multistep methods (Aguilar *et al.*, 2010; VanRaden and Wright, 2013; Legarra *et al.*, 2014). This approach needs the inverse of a dense, genomic relationship matrix ($\boldsymbol{G}^{-1}$; VanRaden, 2008). Therefore, the number of genotyped animals can be a limiting factor in applying ssGBLUP to a population with a large number of genotyped animals.

The “algorithm for proven and young animals” (APY; Misztal *et al.*, 2014) provides a sparser $\boldsymbol{G}^{-1}$ ($\boldsymbol{G}_{APY}^{-1}$). With this algorithm, genotyped animals are divided into two groups: “base” and “non-base” animals. Computing cost and storage size will decrease if many animals are defined as “non-base”. The $\boldsymbol{G}_{APY}^{-1}$ with arbitrary 10,000 “base” animals provided similar genomic enhanced breeding values (**GEBV**) to genomic evaluations from $\boldsymbol{G}^{-1}$ with less computing time and memory requirement (Fragomeni *et al.*, 2015). Validation reliabilities of GEBV with $\boldsymbol{G}_{APY}^{-1}$ in dairy populations have not been discussed. Also, an efficient implementation of APY with a large number of genotyped animals has not been presented.

The objectives of this study were to develop an efficient implementation of $\boldsymbol{G}_{APY}^{-1}$ and to validate genomic predictions for young genotyped bulls in final score for US Holsteins. We also showed validation reliabilities in genomic predictions in the US Jersey population.

Materials and Methods

Computations

We set up the $\boldsymbol{G}_{APY}^{-1}$ using formulas shown by Fragomeni *et al* (2015):

$$\boldsymbol{G}_{APY}^{-1} = \begin{bmatrix} \boldsymbol{G}_{bb}^{-1} + \boldsymbol{G}_{bb}^{-1}\boldsymbol{G}_{bc}\boldsymbol{M}_{cc}^{-1}\boldsymbol{G}'_{bc}\boldsymbol{G}_{bb}^{-1} & -\boldsymbol{G}_{bb}^{-1}\boldsymbol{G}_{bc}\boldsymbol{M}_{cc}^{-1} \\ -\boldsymbol{M}_{cc}^{-1}\boldsymbol{G}'_{bc}\boldsymbol{G}_{bb}^{-1} & \boldsymbol{M}_{cc}^{-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}_{APY}^{bb} & \boldsymbol{G}_{APY}^{bc} \\ \boldsymbol{G}_{APY}^{cb} & \boldsymbol{M}_{cc}^{-1} \end{bmatrix}$$

and

$$\boldsymbol{M}_{cc}^{-1} = \boldsymbol{diag}\{\, \boldsymbol{g}_{ii} - \boldsymbol{g}'_{bi}\boldsymbol{G}_{bb}^{-1}\boldsymbol{g}_{bi} \,\}$$

where $\boldsymbol{G}$ is a genomic relationship matrix, the subscript *b* refers to “base” animals, the subscript *c* refers to “non-base” animals, $\boldsymbol{g}_{ii}$ is diagonal elements in $\boldsymbol{G}$ for “non-base” animal *i*, and $\boldsymbol{g}_{bi}$ is the *i*-th column in $\boldsymbol{G}_{bc}$. The matrix $\boldsymbol{G}_{APY}^{-1}$ was stored as a combination of matrices ($\boldsymbol{G}_{APY}^{bb}$ and $\boldsymbol{G}_{APY}^{bc}$) and a vector ($\boldsymbol{M}_{cc}^{-1}$).

We did not explicitly calculate an inverse of the numerator relationship matrix for genotyped animals ($\boldsymbol{A}_{22}^{-1}$). When mixed model equations are solved with preconditioned conjugate gradient (PCG), only a product of this inverse and a vector, say $\boldsymbol{q}$, is required in each round. Strandén and Mänysaari (2014) showed:

$$\boldsymbol{A}_{22}^{-1}\boldsymbol{q} = [\boldsymbol{A}^{22} - \boldsymbol{A}^{21}(\boldsymbol{A}^{11})^{-1}\boldsymbol{A}^{12}]\boldsymbol{q},$$

where $\boldsymbol{A}^{11}$, $\boldsymbol{A}^{21}$, and $\boldsymbol{A}^{11}$ are sparse submatrices of $\boldsymbol{A}^{-1}$. The product $\boldsymbol{A}_{22}^{-1}\boldsymbol{q}$ was calculated with sparse submatrices.

We used the BLUP90IOD2 program (http://nce.ads.uga.edu/wiki/BLUPmanual) to solve mixed model equations with the PCG algorithm. Dense matrix multiplications in computing $\boldsymbol{G}_{APY}^{-1}$ were performed using a multi-threaded version of the Intel Math Kernel Library (Intel Corporation, Santa Clara, CA). All the analyses were performed on a computer running Linux (x86_64) with Intel Xeon CPU (3.0GHz) processors with 24 cores.

Validation studies

Data

We show the description of data used in this study in Table 1. We used final score from Holstein cows classified up to March, 2014. Genotypes on 60,671 SNP markers were available for 569,404 animals (*n*). These data were referred as the full data set. A truncated data set used for validation contained phenotypes from cows classified in 2009 or earlier.

Table 1. Numbers of phenotypes, recorded cows, pedigree animals, and genotypes in full and truncated data sets for Holsteins.

Data	Number
Full data set	
Phenotypes	11,626,576
Recorded cows	7,093,380
Pedigrees	10,710,380
Genotypes	569,404
Truncated data set	
Phenotypes	10,671,898
Recorded cows	6,384,859

Definitions of “base” animals

We defined 8 “base” groups for Holsteins: genotyped bulls with at least 1 classified daughters up to 2009 (**Base09K**; N = 9,406), the bulls included in Base09K and their dams genotyped and classified up to 2009 (**Base10K**; N = 10,458), the animals included in Base10K, and genotyped and classified cows born up to 2009 (**Base17K**; N = 16,828), and randomly sampled 5,000 (**Rand05K**), 10,000 (**Rand10K**), 15,000 (**Rand15K**), 20,000 (**Rand20K**), and 30,000 (**Rand30K**) animals from a group of 77,066 genotyped animals born in 2009 or earlier. The sampling was replicated 3 times.

Models

A single-trait ssGBLUP model was employed to predict GEBV with the linear animal model described by Tsuruta *et al*. (2002). The mixed model equations included the inverse of the realized relationship matrix (**H**):

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

where τ and ω are constants to reduce bias in GEBV (Misztal *et al.*, 2010). We used τ = 1.0 and ω = 0.9 in this study.

Validation

We defined the "predicted bulls" as young genotyped bulls which had no daughters classified in the truncated data but had at least 30 daughters classified in the full data (N = 2,948).

Daughter deviations (VanRaden and Wiggans, 1991) for the predicted bulls (**DD2014**) were calculated from the full data set without genomic information. Genomic predictions for the predicted bulls (**GEBV2009**) were calculated using the truncated data set. Parent average (**PA2009**) were also calculated without genomic information with the truncated data.

A linear regression analysis was conducted for each combination of DD2014 with genomic predictions (or parent averages) for predicted bulls. The coefficient of determination (R^2) and the regression coefficient (b_1) of DD2014 on genomic predictions were calculated to assess validation reliability and bias, respectively.

Results and Discussion

Table 2 shows R^2 and b_1 of DD2014 on PA2014 and GEBV2009 from various $\mathbf{G}_{\mathrm{APY}}^{-1}$ for the predicted bulls. Genomic predictions always had greater R^2 and b_1 than PA2009. The R^2 and b_1 were almost consistent over the definitions of "base" animals. For randomly sampled "base" animals, R^2 and b_1 were very consistent over replicates. We needed 10,000 or more base animals to achieve the highest R^2. The validation reliabilities in ssGBLUP were greater than 0.40 reported by Tsuruta *et al.* (2013) for 1,851 young bulls in the US Holsteins with 39,741 genotyped animals.

Table 2. Coefficient of determination (R^2) and regression coefficient (b_1) of DD2014 on PA2009 and GEBV2009 for predicted Holstein bulls with at least 30 daughters classified in 2014; average R^2 and b_1 over 3 replicates are shown for random sampled "base" animals.

Prediction	"base" animals	R^2	b_1
PA2009		0.25	0.63
GEBV2009	Base09K	0.44	0.82
	Base10K	0.45	0.82
	Base17K	0.45	0.83
	Rand05K	0.39	0.83
	Rand10K	0.44	0.83
	Rand15K	0.44	0.83
	Rand20K	0.44	0.82
	Rand30K	0.44	0.82

Table 3 shows brief results from validation studies for 305-d milk yield in the US Jersey population (see our presentation for details at http://www.interbull.org/ib/orlando_presentations). Single-step GBLUP resulted in very similar R^2 and b_1 in genomic predictions compared to the multistep method. We observed almost no differences in R^2 and b_1 between $\mathbf{G}^{-1}$ and $\mathbf{G}_{\mathrm{APY}}^{-1}$ in ssGBLUP.

Table 3. Coefficient of determination (R^2) and regression coefficient (b_1) of DD2014 on traditional PTA, multistep GPTA, and single-step GPTA in 2010 for predicted bulls (N = 457) with EBV with at least 75% reliability in 2014 in the US Jersey with 75,053 genotypes.

Prediction	"base" animals	R^2	b_1
Traditional PTA		0.40	0.78
Multistep GPTA[a)]		0.54	0.89
ssGBLUP $\mathbf{G}^{-1}$		0.56	0.84
ssGBLUP $\mathbf{G}_{\mathrm{APY}}^{-1}$	Bulls[b)]	0.55	0.84
	Bulls[c)]	0.56	0.84
	Rand10K	0.55	0.84
	Rand15K	0.56	0.84

a) All tests predicted 482 validation bulls that had no daughters in 2010; b) Old bulls with at least 1 progeny (N = 10,677); c) All bulls with at least 1 progeny (N = 15,960).

Table 4 shows wall-clock time for setting-up $\mathbf{G}_{\mathrm{APY}}^{-1}$ and one iteration in PCG as well as required memory to calculate and store $\mathbf{G}_{\mathrm{APY}}^{-1}$ for a replicate from Rand10K and Rand30K in

US Holsteins. We needed only 7 minutes to prepare the submatrices for $\mathbf{A}_{22}^{-1}$. The maximum memory requirement for $\mathbf{G}_{\mathrm{APY}}^{-1}$ was 151 GB, which can be handled with recent computers. The maximum number of rounds to convergence was 1,329 observed in Rand30K.

Table 4. Wall-clock time for setting-up $\mathbf{G}_{\mathrm{APY}}^{-1}$ and one iteration in PCG, and required memory in Rand10K and Rand30K for 569,404.

	Rand10K	Rand30K
Wall-clock time		
Setting up $\mathbf{G}_{\mathrm{APY}}^{-1}$	1 h 17 m	2 h 45 m
An iteration in PCG	11.7 s	16.5 s
Required memory		
Storage for $\mathbf{G}_{\mathrm{APY}}^{-1}$	42 GB	127 GB
Other	14 GB	24 GB

Our implementation will be capable of running genomic evaluations with more than 570 thousand genotypes. Assume that we have 2 million genotyped animals, 10,000 as "base" animals, and the same number of markers and ancestors to this study. The computing cost for $\mathbf{G}_{\mathrm{APY}}^{-1}$ is proportional to the number of genotyped animals and the storage cost is also the same. The total storage will be 183 GB and the computing time for $\mathbf{G}_{\mathrm{APY}}^{-1}$ will be 4.5 hours. Based on the current timing in Rand10K, a negligible time for $\mathbf{A}_{22}^{-1}$ and 4 more seconds in one PCG-round are expected. If we need 1,000 rounds in PCG, the total computing time for the evaluation will be 10.4 hours. Faster computers can reduce the time.

Conclusions

We conclude that 10,000 or more "base" animals provide accurate genomic predictions in terms of validation reliability. The choice of "base" animals is arbitrary for $\mathbf{G}_{\mathrm{APY}}^{-1}$. Single-step GBLUP with $\mathbf{G}_{\mathrm{APY}}^{-1}$ is computationally applicable to a population with a large number of genotyped animals.

Acknowledgements

We thank Paul VanRaden, Melvin Tooker, and George Wiggans at AGIL, USDA for providing a computing environment and GPTA for Jerseys. Analyses for Jerseys were conducted when the first author stayed at AGIL. This research was primarily supported by grants from Holstein Association USA (Brattleboro, VT) and the U.S. Department of Agriculture's National Institute of Food and Agriculture (Agriculture and Food Research Initiative competitive grant 2015-67015-22936).

References

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

Fragomeni, B.O., Lourenco, D.A.L., Tsuruta, S., Masuda, Y., Aguilar, I., Legarra, A., Lawlor, T.J. & Misztal, I. 2015. Hot topic: use of genomic recursions in single-step genomic BLUP with a large number of genotypes. *Journal of Dairy Science 98*, 4090-4094.

Legarra, A., Christensen, O.F., Aguilar, I. & Misztal, I. 2014. Single Step, a general approach for genomic selection. *Livestock Science 166*, 54-65.

Misztal, I., Aguilar, I., Legarra, A. & Lawlor, T.J. 2010. Choice of parameters for single-step genomic evaluation for type. *Journal of Dairy Science*. *93(Suppl. 1),* p. 533. (Abstr.)

Misztal, I., Legarra, A. & Aguilar, I. 2014. Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science 97*, 3943–3952.

Strandén, I. & Mäntysaari, E.A. 2014. Comparison of some equivalent equations to solve single-step GBLUP. In: *Proceedings of the 10th World Congress on*

Genetics Applied to Livestock Production. Vancouver (Canada). Aug. 17–22, Comm. 069.

Tsuruta, S., Misztal, I., Klei, L. & Lawlor, T.J. 2002. Analysis of age-specific predicted transmitting abilities for final scores in Holsteins with a random regression model. *Journal of Dairy Science 85*, 1324–1330.

Tsuruta, S., Misztal, I. & Lawlor, T.J. 2013. Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. *Journal of Dairy Science 96,* 3332-3335.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science 91*, 4414–4423.

VanRaden, P.M. & Wiggans, G.R. 1991. Derivation, calculation, and use of national animal model information. *Journal of Dairy Science 74,* 2737-2746.

VanRaden, P.M. & Wright, J.R. 2013. Measuring genomic pre-selection in theory and in practice. *Interbull Bulletin 47,* 147-150.