Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals

Y. Masuda, I. Misztal, S. Tsuruta, Andres Legarra, I. Aguilar, D.A.L. Lourenco, B.O. Fragomeni, T.J. Lawlor

▶ **To cite this version:**

Y. Masuda, I. Misztal, S. Tsuruta, Andres Legarra, I. Aguilar, et al.. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. Journal of Dairy Science, 2016, 99 (3), pp.1968-1974. 10.3168/jds.2015-10540 . hal-02635004

HAL Id: hal-02635004
https://hal.inrae.fr/hal-02635004

Submitted on 27 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

J. Dairy Sci. 99:1968–1974
http://dx.doi.org/10.3168/jds.2015-10540
© 2016, THE AUTHORS. Published by FASS and Elsevier Inc. on behalf of the American Dairy Science Association®. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals

Y. Masuda,*[1] I. Misztal,* S. Tsuruta,* A. Legarra,† I. Aguilar,‡ D. A. L. Lourenco,* B. O. Fragomeni,* and T. J. Lawlor§
*Department of Animal and Dairy Science, University of Georgia, Athens 30602
†Institut National de la Recherche Agronomique, UMR1388 GenPhySE, 31326 Castanet Tolosan, France
‡Instituto Nacional de Investigación Agropecuaria, Canelones, Uruguay 90200
§Holstein Association USA Inc., Brattleboro, VT 05301

ABSTRACT

The objectives of this study were to develop and evaluate an efficient implementation in the computation of the inverse of genomic relationship matrix with the recursion algorithm, called the algorithm for proven and young (APY), in single-step genomic BLUP. We validated genomic predictions for young bulls with more than 500,000 genotyped animals in final score for US Holsteins. Phenotypic data included 11,626,576 final scores on 7,093,380 US Holstein cows, and genotypes were available for 569,404 animals. Daughter deviations for young bulls with no classified daughters in 2009, but at least 30 classified daughters in 2014 were computed using all the phenotypic data. Genomic predictions for the same bulls were calculated with single-step genomic BLUP using phenotypes up to 2009. We calculated the inverse of the genomic relationship matrix $\left(\mathbf{G}^{-1}_{\text{APY}}\right)$ based on a direct inversion of genomic relationship matrix on a small subset of genotyped animals (core animals) and extended that information to non-core animals by recursion. We tested several sets of core animals including 9,406 bulls with at least 1 classified daughter, 9,406 bulls and 1,052 classified dams of bulls, 9,406 bulls and 7,422 classified cows, and random samples of 5,000 to 30,000 animals. Validation reliability was assessed by the coefficient of determination from regression of daughter deviation on genomic predictions for the predicted young bulls. The reliabilities were 0.39 with 5,000 randomly chosen core animals, 0.45 with the 9,406 bulls, and 7,422 cows as core animals, and 0.44 with the remaining sets. With phenotypes truncated in 2009 and the preconditioned conjugate gradient to solve mixed model equations, the number of rounds to convergence for core animals defined by bulls was 1,343; defined by bulls and cows, 2,066; and defined by 10,000 random animals, at most 1,629. With complete phenotype data, the number of rounds decreased to 858, 1,299, and at most 1,092, respectively. Setting up $\mathbf{G}^{-1}_{\text{APY}}$ for 569,404 genotyped animals with 10,000 core animals took 1.3 h and 57 GB of memory. The validation reliability with APY reaches a plateau when the number of core animals is at least 10,000. Predictions with APY have little differences in reliability among definitions of core animals. Single-step genomic BLUP with APY is applicable to millions of genotyped animals.
Key words: final score, genomic relationship matrix, genomic evaluation

INTRODUCTION

Single-step genomic BLUP (**ssGBLUP**) is a tool for genomic evaluations (Aguilar et al., 2010; Christensen and Lund, 2010). The method has numerous advantages over multistep methods: simplicity, avoidance of double counting, and resistance to biased prediction caused by preselection of young animals (Patry and Ducrocq, 2011; Vitezica et al., 2011; VanRaden and Wright, 2013; Legarra et al., 2014). In this method, mixed model equations contain the inverse of a genomic relationship matrix ($\mathbf{G}$). The cost of inversion and the storage of the inverse are proportional to the cubic and quadratic of the number of genotyped animals (n_g), respectively. Therefore, the large computing cost is a limiting factor in an application of ssGBLUP to an actual population with a large number of genotyped animals such as the US Holsteins with more than 900,000 genotyped animals (https://www.cdcb.us/Genotype/cur_density.html).

Received October 19, 2015.
Accepted December 1, 2015.
[1]Corresponding author: yutaka@uga.edu

Misztal et al. (2014) suggested calculating a sparse inverse of $\mathbf{G}$ with an algorithm for proven and young animals (**APY**) based on recursive equations. We will refer to the inverse from this algorithm as the $\mathbf{G}^{-1}_{\text{APY}}$ compared with the regular $\mathbf{G}^{-1}$. Initial studies used the term proven for the starting animals and the term young for the younger animals in the genomic recursions, thus the name "algorithm for proven and young." Subsequent research (Fragomeni et al., 2015) showed that it is not necessary to order animals by age, whereby a smaller number of core animals can be used in the direct inversion and recursion used on the noncore animals.

In the algorithm in this study, genotyped animals were divided into 2 groups: core and noncore, which were labeled as proven and young, respectively, in the earlier studies. In the $\mathbf{G}^{-1}_{\text{APY}}$, off-diagonal elements corresponding to relationships among noncore animals are set to be 0, which can reduce the amount of computations and memory. Fragomeni et al. (2015) reported that the computing cost and storage size for $\mathbf{G}^{-1}_{\text{APY}}$ could be only 0.3 and 8% of those required in the regular $\mathbf{G}^{-1}$, respectively, when n_g was 500,000 and the number of core animals was 20,000.

With at least 10,000 core animals, the $\mathbf{G}^{-1}_{\text{APY}}$ provided genomic EBV (**GEBV**) that were very similar to genomic evaluations from the regular $\mathbf{G}^{-1}$ for US Holsteins (Fragomeni et al., 2015) and American Angus (Lourenco et al., 2015). The previous studies focused on the feasibility of APY, and an efficient computing was not a priority. The objectives of this study were (1) to develop an efficient implementation in the computation of $\mathbf{G}^{-1}_{\text{APY}}$ and evaluate the computational costs and (2) to validate genomic predictions for young bulls using more than 500,000 genotyped animals in final score for US Holsteins.

MATERIALS AND METHODS

Animal Care and Use Committee approval was not obtained for this study because no animals were used.

Computational Strategies

The genomic relationship matrix is typically calculated with the first method described in VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_j\left(1-p_j\right)} \text{ and } \mathbf{Z} = \left(\mathbf{M}-\mathbf{P}\right),$$

where p_j is the allele frequency of the second allele at locus j, $\mathbf{M}$ is a matrix containing marker genotypes, and $\mathbf{P}$ is a matrix containing $2p_j$. The allele frequencies were calculated from the current genotyped animals. We divided $\mathbf{G}$ into 4 submatrices as

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{G}_{nn} \end{bmatrix},$$

where subscript c refers to a group of core animals and subscript n refers to a group of noncore animals. The matrix $\mathbf{G}$ was blended with $\mathbf{A}_{22}$ as $0.95\mathbf{G} + 0.05\mathbf{A}_{22}$ to guarantee its nonsingularity, where $\mathbf{A}_{22}$ is a numerator relationship matrix for genotyped animals. Although each column of $\mathbf{A}_{22}$ were fully calculated using a method described by Aguilar et al. (2011), only the elements corresponding to $\mathbf{G}_{cc}$, $\mathbf{G}_{cn}$, and the diagonals of $\mathbf{G}_{nn}$ were added to $\mathbf{G}$. Then, the blended $\mathbf{G}$ was scaled to satisfy AvgDiag(scaled $\mathbf{G}$) = AvgDiag($\mathbf{A}_{22}$) and AvgOff(scaled $\mathbf{G}$) = AvgOff($\mathbf{A}_{22}$), where AvgDiag($\mathbf{X}$) and AvgOff($\mathbf{X}$) are averages of diagonal and off-diagonal elements of a square matrix $\mathbf{X}$, respectively (Chen et al., 2011; Vitezica et al., 2011). The scaling was expected to reduce the biases in GEBV for young animals (Chen et al., 2011; Vitezica et al., 2011).

Instead of explicitly inverting full $\mathbf{G}$, we set up the $\mathbf{G}^{-1}_{\text{APY}}$ with formulas as in Fragomeni et al. (2015):

$$\mathbf{G}^{-1}_{\text{APY}} = \begin{bmatrix} \mathbf{G}^{-1}_{cc} + \mathbf{G}^{-1}_{cc}\mathbf{G}_{cn}\mathbf{M}^{-1}_{nn}\mathbf{G}'_{cn}\mathbf{G}^{-1}_{cc} & -\mathbf{G}^{-1}_{cc}\mathbf{G}_{cn}\mathbf{M}^{-1}_{nn} \\ -\mathbf{M}^{-1}_{nn}\mathbf{G}'_{cn}\mathbf{G}^{-1}_{nn} & \mathbf{M}^{-1}_{nn} \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{G}^{cc}_{\text{APY}} & \mathbf{G}^{cn}_{\text{APY}} \\ \mathbf{G}^{cn}_{\text{APY}}{}' & \mathbf{M}^{-1}_{nn} \end{bmatrix} \text{ and}$$

$$\mathbf{M}_{nn} = \text{diag}\left\{g_{ii} - \mathbf{g}'_{ci}\mathbf{G}^{-1}_{cc}\mathbf{g}_{ci}\right\},$$

where g_{ii} is the ith diagonal element in $\mathbf{G}_{nn}$, $\mathbf{g}_{ci}$ is the ith column in $\mathbf{G}_{cn}$, and $\mathbf{M}_{nn}$ is a diagonal matrix. Misztal et al. (2014) suggested APY as an algorithm that would only ignore information from nonrelevant animals, such as young animals. Surprisingly, Lourenco et al. (2014) and Fragomeni et al. (2015) observed that a large set of core animals gives $\mathbf{G}^{-1}_{\text{APY}}$ with good accuracy. These results imply parts of $\mathbf{G}^{-1}$ that have no contributions to the accuracy are set to 0 in $\mathbf{G}^{-1}_{\text{APY}}$ (Misztal, 2016). The following steps were implemented to facilitate efficient computations:

1. Genotypes were stored in compressed form to save memory. The value of each locus was coded with 2 bits as: 00 = homozygote, 01 = heterozygote, 10 = another homozygote, and 11 = missing. With this encoding, 16 markers were packed into a 4-byte integer variable. For exam-

ple, 60,000 marker-genotypes were packed into 15,000 bytes (14.6 KB) per animal. The packed genotypes were expanded as needed. Both packing and unpacking were implemented in parallel.
2. Whereas $\mathbf{G}_{cc}$ and $\mathbf{G}_{cn}$ were fully computed and stored as dense matrices, only the diagonals of $\mathbf{G}_{nn}$ were computed and stored as a vector. These 3 matrices were blended with $\mathbf{A}_{22}$ and scaled with the methods described above.
3. We calculated $\mathbf{G}_{cc}^{-1}$ by updating $\mathbf{G}_{cc}$ with the Linear Algebra Package (**LAPACK**; Aguilar et al., 2011). Then, $-\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}$ was stored in temporary memory and the diagonals of $\mathbf{G}_{nn}$ were replaced with the diagonals of $\mathbf{M}_{nn}^{-1}$. Finally, $\mathbf{G}_{cc}$ and $\mathbf{G}_{cn}$ were replaced with $\mathbf{G}_{\text{APY}}^{cc}$ and $\mathbf{G}_{\text{APY}}^{cn}$, respectively. The Basic Linear Algebra Subprograms (**BLAS**) were used for the dense matrix multiplications (Aguilar et al., 2011).

Single-step GBLUP also requires $\mathbf{A}_{22}^{-1}$ (Aguilar et al., 2010; Christensen and Lund, 2010). This matrix was dense and could not be stored in memory. When mixed model equations are solved with the preconditioned conjugate gradient (**PCG**) algorithm, only a product of this matrix and a vector, say $\mathbf{q}$, is required in each round. The product, $\mathbf{A}_{22}^{-1}\mathbf{q}$, was calculated with an equation shown by Strandén and Mäntysaari (2014):

$$\mathbf{A}_{22}^{-1}\mathbf{q} = \left[\mathbf{A}^{22} - \left(\mathbf{A}^{12}\right)' \left(\mathbf{A}^{11}\right)^{-1} \mathbf{A}^{12}\right]\mathbf{q},$$

where $\mathbf{A}^{22}$, $\mathbf{A}^{12}$, and $\mathbf{A}^{11}$ are sparse submatrices of $\mathbf{A}^{-1}$ and subscripts 1 and 2 refer to groups of nongenotyped and genotyped animals, respectively. We set up $\mathbf{A}^{22}$, $\mathbf{A}^{12}$, and $\mathbf{A}^{11}$ using the rapid method (Henderson, 1976; Quaas, 1976) as sparse matrices stored in memory at the preparation phase. In each PCG round, $\mathbf{A}_{22}^{-1}\mathbf{q}$ was calculated with a set of sparse operations: $\mathbf{t} = \mathbf{A}^{22}\mathbf{q}$, $\mathbf{x} = \mathbf{A}^{12}\mathbf{q}$, $\mathbf{y} = (\mathbf{A}^{11})^{-1}\mathbf{x}$, $\mathbf{z} = (\mathbf{A}^{12})'\mathbf{y}$, and $\mathbf{A}_{22}^{-1}\mathbf{q} = \mathbf{t} - \mathbf{z}$, where $\mathbf{t}$, $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ are temporary vectors. Because $\mathbf{y}$ was essentially the solution of sparse equations, we needed the Cholesky factor of $\mathbf{A}^{11}$. The sparse factorization was performed using the YAMS sparse package (Masuda et al., 2014, 2015). In practice, and because it gives strictly the same result, the expression above only considered genotyped animals (in matrix $\mathbf{A}_{22}$) and their nongenotyped ancestors (in matrix $\mathbf{A}_{11}$), which greatly reduces computations.

We incorporated our implementation into the BLUP90IOD2 program (http://nce.ads.uga.edu/wiki/doku.php?id = application_programs), which solves mixed model equations using the PCG algorithm (Tsuruta et al., 2001). The program stopped when the convergence criterion (the squared ratio of the norm of residual and right-hand-side vectors; Tsuruta et al., 2001) was less than 10^{-15}. The program was compiled with the Intel Fortran Compiler 14.0 (Intel Corporation, Santa Clara, CA). We used multi-threaded version of BLAS and LAPACK in the Intel Math Kernel Library 11.0 (Intel Corporation). All the analyses were performed on a computer running Linux (x86_64) with Intel Xeon E7–8857 CPU (3.0 GHz) processors with 24 computing cores.

Data

Phenotypic data for final score and pedigree information were provided by Holstein Association USA Inc. (Brattleboro, VT). The phenotypic data included 11,626,576 records from 7,093,380 cows classified up to March 2014, and pedigree data included 10,710,380 animals. Genotypes for 60,671 SNP markers were available for 569,404 animals. We will refer to these data as the full data set. A truncated data set used for validation contained only 10,671,898 phenotypes from cows classified in 2009 or earlier.

Model

A single-trait ssGBLUP model was employed to predict GEBV. The mixed model equations (Tsuruta et al., 2002) included

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

where τ and ω are scaling factors to reduce bias in GEBV (Misztal et al., 2010, 2013). We used $\tau = 1.0$ and $\omega = 0.9$ in this study.

Validation

We defined the predicted bulls as young genotyped bulls that had no daughters classified in the truncated data (i.e., by the end of 2009) but had at least 30 daughters classified by March 2014 (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 with different definitions of the core animals for $\mathbf{G}_{\text{APY}}^{-1}$. The following definitions were considered: genotyped bulls with at least 1 classified daughter up to 2009 (**Core09K**; n = 9,406), the bulls included in Core09K and their dams genotyped and classified up to

Table 1. Wall-clock time[1] and required memory for preparation of an algorithm for proven and young (APY)-inverse $\left(\mathbf{G}_{\text{APY}}^{-1}\right)$, the computation for components of the inverse of a numerator relationship matrix $\left(\mathbf{A}_{22}^{-1}\right)$, and an iteration in preconditioned conjugate gradient for randomly sampled 5,000, 10,000, 15,000, 20,000, and 30,000 core animals with 569,404 genotyped animals

	Number of core animals				
Item	5,000	10,000	15,000	20,000	30,000
Wall-clock time					
Setting-up $\mathbf{G}_{\text{APY}}^{-1}$					
Load markers (min)	18	18	18	18	18
Compute $\mathbf{G}_{\text{APY}}$ (min)	15	26	37	48	71
Blend with $\mathbf{A}_{22}$ (min)	26	26	26	25	25
Compute $\mathbf{G}_{\text{APY}}^{-1}$ (min)	2	7	14	24	51
Total (h:min)	1:1	1:17	1:35	1:55	2:45
Setting up $\mathbf{A}_{22}^{-1}$					
Preparation (min)	7	7	8	7	7
Iteration					
Per round (s)	10.2	11.7	12.2	13.7	16.5
Required memory (GB)					
Packed markers	8.0	8.0	8.0	8.0	8.0
Working area[2]	4.2	6.4	8.7	10.9	15.5
Storage for $\mathbf{G}_{\text{APY}}^{-1}$	21.2	42.4	63.6	84.9	127.3

[1]Using central processing units (3.0 GHz) with 24 computing cores.

[2]Peak amount required for setting up $\mathbf{G}_{\text{APY}}$ and $\mathbf{G}_{\text{APY}}^{-1}$.

2009 (**Core10K**; n = 10,458), the animals included in Core10K and genotyped and classified cows born up to 2009 (**Core17K**; n = 16,828), and randomly sampled 5,000 (**Rand05K**), 10,000 (**Rand10K**), 15,000 (**Rand15K**), 20,000 (**Rand20K**), and 30,000 (**Rand30K**) animals from the 77,066 genotyped animals born in 2009 or earlier. The random sampling was replicated 3 times. Parent averages (**PA2009**) for the predicted bulls were also calculated with the truncated data set using a traditional animal model.

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 regression coefficient (b_1) of DD2014 on GEBV2009 (or PA2009) were calculated as a validation reliability and a bias indicator, respectively.

RESULTS AND DISCUSSION

Table 1 shows wall-clock time for setting up $\mathbf{G}_{\text{APY}}^{-1}$ and one round in PCG as well as required memory to calculate and store $\mathbf{G}_{\text{APY}}^{-1}$ for a replicate from Rand05K, Rand10K, Rand15K, Rand20K, and Rand30K. The most computationally demanding scenario was Rand30K; setting up $\mathbf{G}_{\text{APY}}^{-1}$ finished within 3 h; one round of PCG required 16.5 s, corresponding to 4.5 h for 1,000 rounds. The maximum memory requirement for $\mathbf{G}_{\text{APY}}^{-1}$ was 130 GB, which can be handled with a small-size server. For the components of $\mathbf{A}_{22}^{-1}$, the computations and the storage were negligible.

The computing time for setting-up $\mathbf{G}_{\text{APY}}^{-1}$ is predictable based on the results from fewer numbers of core animals. Loading markers from a text file to memory took 18 min. Computing time both in packing the marker genotypes and expanding the packed code was only 2 min on average. Running time for setting up partial $\mathbf{G}$ and the computations for $\mathbf{G}_{\text{APY}}^{-1}$ were approximately proportional to n_c and n_c^2, respectively, where n_c is the number of core animals. The results agreed with the theoretical evaluation of computing costs (Misztal et al., 2014; Fragomeni et al., 2015).

Table 2 shows the number of rounds to convergence in PCG. We needed more rounds when more core animals were included, as reported by Fragomeni et al. (2015). Koivula et al. (2015) reported that the PCG algorithm failed to converge within 5,000 rounds in ssGBLUP with random regressions when the matrix $\mathbf{G}$ was scaled. We used the same scaling method and all the PCG algorithm converged within 3,400 rounds in all the equations tested. Validation data contained many descendant animals without phenotypes, and the data structure caused the poor convergence rate (Legarra et al., 2014). When the full data were used, we needed fewer rounds to reach the same convergence criterion (Table 2).

Table 3 shows R^2 and b_1 (squared accuracy and bias) of DD2014 on PA2009 and GEBV2009 from various $\mathbf{G}_{\text{APY}}^{-1}$. Genomic predictions always had greater accuracy and less bias than PA2009. The R^2 and b_1 were very similar across different sets of core animals. For ran-

Table 2. Rounds to convergence in preconditioned conjugate gradient for different definitions of core animals with truncated phenotypes, all the pedigrees, and all the genotyped animals[1] (truncated data) and with the nontruncated data (full data)

Model	Definition of core animals[2]	Rounds to convergence	
		Truncated data	Full data
Traditional BLUP		699	571
Single-step GBLUP	Core09K	1,343	858
	Core10K	1,502	911
	Core17K	2,066	1,299
	Rand05K[3]	1,049–1,159	671–694
	Rand10K[3]	1,581–1,629	1,027–1,092
	Rand15K[3]	1,952–2,094	1,260–1,419
	Rand20K[3]	2,327–2,505	1,491–1,670
	Rand30K[3]	2,874–3,329	1,870–1,976

[1]Including 569,404 genotyped animals.

[2]Core09K = genotyped bulls with at least 1 classified daughter up to 2009 (n = 9,406); Core10K = Core09K + their dams genotyped and classified up to 2009 (n = 10,458); Core17K = Core10K + genotyped and classified cows born up to 2009 (n = 16,828); Rand05K, Rand10K, Rand15K, Rand20K, and Rand30K = randomly sampled 5,000, 10,000, 15,000, 20,000, and 30,000 animals, respectively. All the core animals were from the 77,066 genotyped animals born in 2009 or earlier.

[3]Ranges over 3 replicates are shown.

domly sampled core animals, R^2 and b_1 were almost consistent over replicates. The greatest R^2 was achieved with 10,000 or more core animals. Lourenco et al. (2015) reported that GEBV from $\mathbf{G}^{-1}_{\text{APY}}$ with 8,000 core animals accounted for 97% of prediction accuracy from GEBV with $\mathbf{G}^{-1}$ for birth weight in the US Angus population. Figure 1 shows the correlations of GEBV2009 from Rand30K with GEBV from the other definitions of core animals. The correlation approached to 1 as n_c increased. This trend is similar to the results from Fragomeni et al. (2015) who used 100,000 genotypes from the US Holstein population. Over all, our results agree with Fragomeni et al. (2015) who concluded that 10,000 or more core animals provided accurate genomic evaluations and randomly sampled core animals could achieve the correlation of 1 between GEBV estimated either with $\mathbf{G}^{-1}$ or with $\mathbf{G}^{-1}_{\text{APY}}$.

The validation reliabilities (Table 3) were generally greater than the previous reports for final score in the US Holstein population. Tsuruta et al. (2013) reported the validation reliability of 0.40 for 1,851 young bulls to be sires with at least 30 daughters using ssGBLUP with 39,741 genotyped animals. Tsuruta et al. (2014) showed the validation reliability of 0.34 for 2,425 young

Table 3. Coefficients of determination (R^2) and regression coefficients (b_1) of daughter deviations in 2014 (DD2014) on parent average (PA2009) and genomic breeding values (GEBV2009) from algorithm for proven and young-inverses using truncated phenotypes, all the pedigrees, and all the genotyped animals[1] for the predicted young bulls with at least 30 daughters classified in 2014

Sire evaluation	Definition of core animals[2]	Predicted bulls (n = 2,948)	
		R^2	b_1
PA2009		0.25	0.63
GEBV2009	Core09K	0.44	0.82
	Core10K	0.45	0.82
	Core17K	0.45	0.83
	Rand05K[3]	0.39–0.39	0.74–0.75
	Rand10K[3]	0.43–0.44	0.83–0.84
	Rand15K[3]	0.44–0.44	0.83–0.84
	Rand20K[3]	0.44–0.44	0.82–0.83
	Rand30K[3]	0.44–0.44	0.82–0.83

[1]Including 569,404 genotyped animals.

[2]Core09K = genotyped bulls with at least 1 classified daughter up to 2009 (n = 9,406); Core10K = Core09K + their dams genotyped and classified up to 2009 (n = 10,458); Core17K = Core10K + genotyped and classified cows born up to 2009 (n = 16,828); Rand05K, Rand10K, Rand15K, Rand20K, and Rand30K = randomly sampled 5,000, 10,000, 15,000, 20,000, and 30,000 animals, respectively. All the core animals were from the 77,066 genotyped animals born in 2009 or earlier.

[3]Ranges of R^2 and b_1 over 3 replicates are shown.

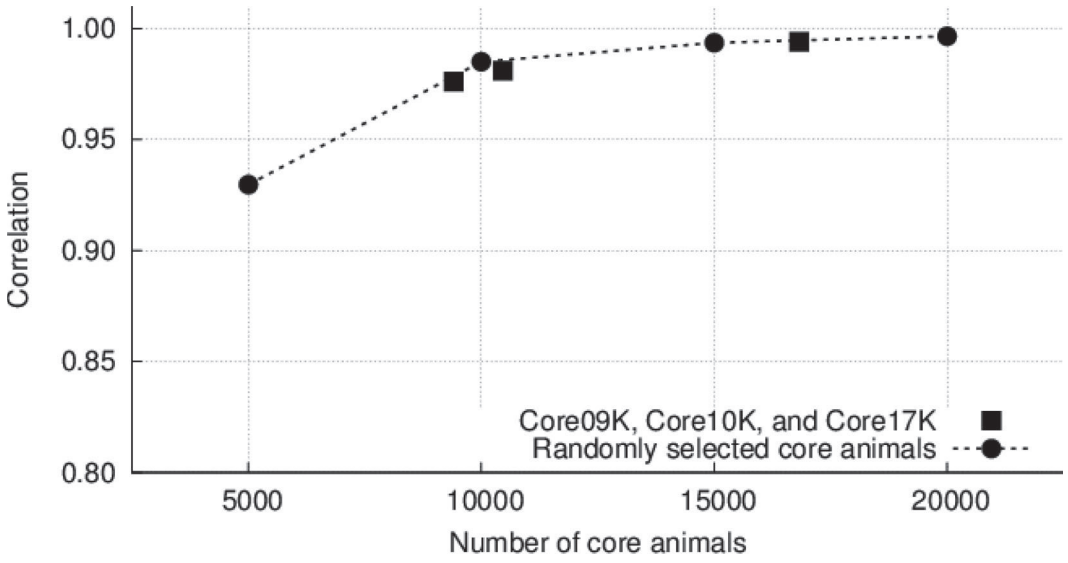

Figure 1. Correlation coefficients between genomic EBV (GEBV2009) from randomly sampled 30,000 core animals (Rand30K) and GEBV2009 from the other definitions of the core animals. Genomic EBV for the predicted young bulls with at least 30 daughters classified in 2014 (n = 2,948) were computed with the algorithm for proven and young-inverse of genomic relationship matrix. Core animals were defined as Core09K = genotyped bulls with at least 1 classified daughter up to 2009 (n = 9,406); Core10K = Core09K + their dams genotyped and classified up to 2009 (n = 10,458); Core17K = Core10K + genotyped and classified cows born up to 2009 (n = 16,828); Rand05K, Rand10K, Rand15K, Rand20K, and Rand30K = randomly sampled 5,000, 10,000, 15,000, 20,000, and 30,000 animals, respectively. For randomly sampled core animals, the correlation coefficient was calculated in each combination of replicates and the average was shown.

bulls to be sires with at least 10 daughters, using ssGBLUP with 34,500 genotyped bulls. Olson et al. (2011) used a multistep method with 8,022 genotyped bulls as reference animals and reported the validation reliability of 0.34 for 2,653 young bulls. The regression coefficients (Table 3) were similar to or slightly less than the previous studies: 0.81 (Tsuruta et al., 2013), 0.85 (Tsuruta et al., 2014), and 0.86 (Olson et al., 2011). The greater reliability and lesser regression coefficient can be related to the larger number of genotyped animals as well as the parameter ω. For example, when we used Core09K with $\omega = 0.7$, R^2 was 0.35, and b_1 was 0.96 (results not shown). Tsuruta et al. (2013) and Koivula et al. (2015) reported that a greater ω resulted in more accurate (greater R^2) although more biased (smaller b_1) predictions for young animals. The ω parameter may partly compensate for incomplete pedigree information (Misztal et al., 2013). Truncation of old pedigrees and phenotypes had a minimal effect on realized accuracies (Lourenco et al., 2014), but an effect of pedigree completeness on the reliability has not been investigated thus far.

Based on the results of this study, we can estimate costs for a genomic evaluation using $\mathbf{G}^{-1}_{\text{APY}}$ involving 2 million genotyped animals. Assuming 10,000 core animals and the formulas in the Appendix, the total storage will be 183 GB (149 GB for the storage of $\mathbf{G}^{-1}_{\text{APY}}$), and the computing time to set up $\mathbf{G}^{-1}_{\text{APY}}$ will be 4.5 h. We expect a negligible time for the preparation of the components of $\mathbf{A}^{-1}_{22}$. The multiplication $\mathbf{G}^{-1}_{\text{APY}}\mathbf{q}$ in PCG will need 9.5 more seconds per iteration based on the current timing in Rand10K, in which the multiplication required 3.8 s per iteration. If we need 1,000 rounds in PCG, the total computing time for the evaluation will be 10.4 h. The $\mathbf{G}^{-1}_{\text{APY}}$ hence removes the computational limitations caused by the number of genotyped animals in ssGBLUP.

ACKNOWLEDGMENTS

This research was primarily supported by grants from Holstein Association USA (Brattleboro, VT) and the USDA's National Institute of Food and Agriculture (Washington, DC; Agriculture and Food Research Initiative competitive grant 2015-67015-22936).

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.

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128:422–428.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci. 89:2673–2679. http://dx.doi.org/10.2527/jas.2010-3555.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.
Fragomeni, B. O., D. A. L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. 2015. Hot topic: Use of genomic recursions in single-step genomic BLUP with a large number of genotypes. J. Dairy Sci. 98:4090–4094.
Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83.
Koivula, M., I. Strandén, J. Pösö, G. P. Aamand, and E. A. Mäntysaari. 2015. Single-step genomic evaluation using multitrait random regression model and test-day data. J. Dairy Sci. 98:2775–2784.
Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livest. Sci. 166:54–65.
Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, T. J. Lawlor, S. Forni, and J. I. Weller. 2014. Are evaluations on young genotyped animals benefiting from the past generations? J. Dairy Sci. 97:3930–3942.
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 BLUP in American Angus. J. Anim. Sci. 93:2653–2662.
Masuda, Y., I. Aguilar, S. Tsuruta, and I. Misztal. 2015. Technical note: Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements. J. Anim. Sci. 93:4670–4674.
Masuda, Y., T. Baba, and M. Suzuki. 2014. Application of supernodal sparse factorization and inversion to the estimation of (co)variance components by residual maximum likelihood. J. Anim. Breed. Genet. 131:227–236.
Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics http://dx.doi.org/10.1534/genetics.115.182089.
Misztal, I., I. Aguilar, A. Legarra, and T. J. Lawlor. 2010. Choice of parameters for single-step genomic evaluation for type. J. Dairy Sci. 93(Suppl. 1):533. (Abstr.)
Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97:3943–3952.
Misztal, I., Z. G. Vitezica, A. Legarra, I. Aguilar, and A. A. Swan. 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130:252–258.
Olson, K. M., P. M. VanRaden, M. E. Tooker, and T. A. Cooper. 2011. Differences among methods to validate genomic evaluations for dairy cattle. J. Dairy Sci. 94:2613–2620.
Patry, C., and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J. Dairy Sci. 94:1011–1020.
Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32:949–953.
Strandén, I., and E. A. Mäntysaari. 2014. Comparison of some equivalent equations to solve single-step GBLUP. No. 069 in Proc. 10th WCGALP, Vancouver, Canada. Accessed Aug. 20, 2015. https://asas.org/docs/default-source/wcgalp-proceedings-oral/069_paper_9344_manuscript_568_0.pdf.
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.
Tsuruta, S., I. Misztal, and T. J. Lawlor. 2013. Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. J. Dairy Sci. 96:3332–3335.
Tsuruta, S., I. Misztal, D. A. L. Lourenco, and T. J. Lawlor. 2014. Assigning unknown parent groups to reduce bias in genomic evaluations of final score in US Holsteins. J. Dairy Sci. 97:5814–5821.
Tsuruta, S., I. Misztal, and I. Strandén. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. J. Anim. Sci. 79:1166–1172.
VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.
VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74:2737–2746.
VanRaden, P. M., and J. R. Wright. 2013. Measuring genomic preselection in theory and in practice. Interbull Bull. 47:147–150.
Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. (Camb.) 93:357–366.

APPENDIX

Future Costs of Storage and Computation in ssGBLUP with APY

We assume that (1) the number of markers (m) is constant, (2) the number of genotyped animals (n_g) is growing, and (3) the number of core animals (n_c) is small and constant (i.e., $n_c \ll n_g$ and $n_n = n_g - n_c \approx n_g$, where n_n is the number of noncore animals). We also assume that newly genotyped animals are all the descendants of the current genotyped animals so the number of ancestors for genotyped animals is also fixed.

The most memory is consumed for the dense blocks in $\mathbf{G}^{-1}_{\text{APY}}$ and the number of elements in the blocks is $n_c \times n_g$, so the cost is $O(n_g)$, where $O(\cdot)$ is the big-O notation representing a theoretical measure of the time or memory needed (https://xlinux.nist.gov/dads/HTML/bigOnotation.html). The storage cost for the markers is also $O(n_g)$. In our implementation, we used temporary memory to store $-\mathbf{G}^{-1}_{cc}\mathbf{G}_{cn}$ and an amount of the temporary memory depends only on n_c (i.e., the storage cost for the temporary memory is constant).

In the computations, the most time-consuming processes are $[\mathbf{G}_{cc}\ \mathbf{G}_{cn}] = \mathbf{Z}_c\left[\mathbf{Z}'_c\ \ \mathbf{Z}'_n\right]$ for $\mathbf{G}$ ($\mathbf{Z}_c$ and $\mathbf{Z}_n$ are submatrices of $\mathbf{Z}$ for core and noncore animals, respectively) and $\mathbf{G}^{-1}_{cc}\mathbf{G}_{cn}\mathbf{M}^{-1}_{nn}\mathbf{G}'_{cn}\mathbf{G}^{-1}_{cc}$ for $\mathbf{G}^{-1}_{\text{APY}}$. The former needs $n_c \times n_g \times m$ operations, and the latter needs $2n_c^2 \times n_n$ operations so that the computing cost is $O(n_g)$ for both $\mathbf{G}$ and $\mathbf{G}^{-1}_{\text{APY}}$. The cost in reading markers is $O(n_g)$. In PCG, a multiplication of $\mathbf{G}^{-1}_{\text{APY}}\mathbf{q}$ needs $n_c^2 + 2n_c \times n_n + n_n$ operations and the cost is $O(n_g)$.

As more genotyped animals are available, the number of nonzero elements increases in $\mathbf{A}^{22}$ but is constant in $\mathbf{A}^{12}$ and $\mathbf{A}^{11}$. Therefore, the future computation cost increases for $\mathbf{A}^{22}\mathbf{q}$ but remains in $(\mathbf{A}^{12})'(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}\mathbf{q}$. The product $\mathbf{A}^{22}\mathbf{q}$ can be directly calculated from a pedigree list (Henderson, 1976; Quaas, 1976), and the cost is $O(n_g)$. The cost for storage of $\mathbf{A}^{22}$ is also $O(n_g)$.