J. Dairy Sci. 98:4090–4094
http://dx.doi.org/10.3168/jds.2014-9125

Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes

B. O. Fragomeni,*[1] D. A. L. Lourenco,* S. Tsuruta,* Y. Masuda,* I. Aguilar,† A. Legarra,‡ T. J. Lawlor,§ and I. Misztal*
*Department of Animal and Dairy Science, University of Georgia, Athens 30602
†Instituto Nacional de Investigacion Agropecuaria, Canelones, 90200, Uruguay
‡INRA, UMR1388 GenePhySE, Castanet Tolosan, 31326, France
§Holstein Association USA Inc., Brattleboro, VT 05302

ABSTRACT

The purpose of this study was to evaluate the accuracy of genomic selection in single-step genomic BLUP (ssGBLUP) when the inverse of the genomic relationship matrix (**G**) is derived by the "algorithm for proven and young animals" (APY). This algorithm implements genomic recursions on a subset of "proven" animals. Only a relationship matrix for animals treated as "proven" needs to be inverted, and the extra costs of adding animals treated as "young" are linear. Analyses involved 10,102,702 final scores on 6,930,618 Holstein cows. Final score, which is a composite of type traits, is popular trait in the United States and was easily available for this study. A total of 100,000 animals with genotypes were used in the analyses and included 23,000 sires (16,000 with >5 progeny), 27,000 cows, and 50,000 young animals. Genomic EBV (GEBV) were calculated with a regular inverse of **G**, and with the **G** inverse approximated by APY. Animals in the proven subset included only sires (23,000), sires + cows (50,000), only cows (27,000), or sires with >5 progeny (16,000). The correlations of GEBV with APY and regular GEBV for young genotyped animals were 0.994, 0.995, 0.992, and 0.992, respectively Later, animals in the proven subset were randomly sampled from all genotyped animals in sets of 2,000, 5,000, 10,000, 15,000, and 20,000; each sample was replicated 4 times. Respective correlations were 0.97 (5,000 sample), 0.98 (10,000 sample), and 0.99 (20,000 sample), with minimal difference between samples of the same size. Genomic EBV with APY were accurate when the number of animals used in the subset is between 10,000 and 20,000, with little difference between the ways of creating the subset. Due to the approximately linear cost of APY, ssGBLUP with APY could support any number of genotyped animals without affecting accuracy.

Key words: single-step method, genomic selection, genomic recursion

INTRODUCTION

Single-step genomic BLUP (**ssGBLUP**; Aguilar et al., 2010; Christensen and Lund, 2010) has emerged as a simple yet accurate tool for genetic evaluations. Its main advantages over multistep methods are simplicity, no double counting, and resistance to preselection bias (Vitezica et al., 2011; VanRaden and Wright, 2013; Legarra et al., 2014). As originally defined, ssGBLUP uses classical BLUP mixed equations extended with the inverse of the genomic (**G**) and pedigree ($\mathbf{A}_{22}$) relationship matrices for genotyped animals. With algorithms as described in Aguilar et al. (2011), the cost of obtaining these matrices is cubic, and currently there is a soft limit of about 150k genotyped animals in the model; however, >600k genotyped animals are available for US Holsteins (https://www.cdcb.us/Genotype/cur_density.html). Several approaches have been proposed to overcome such a limit (Legarra and Ducrocq, 2012; Fernando et al., 2014; Liu et al., 2014) but either they have convergence problems or are expensive and hard to program and use with data and a variety of models such as multiple trait or random regressions.

Faux et al. (2012) attempted to extend the rules used in creation of the numerator relationship matrix to approximate the inverse of **G**. Their method was based on incomplete Cholesky factorization, where only genomic relationships between close relatives were considered. However, the approximation was not accurate enough, and steps proposed to increase that accuracy were expensive.

Recently, Misztal et al. (2014) proposed a method based on genomic recursion, where genomic breeding value (**GBV**) of a new genotyped animal is conditioned on GBV of all the previous genotyped animals. One of their proposed algorithms was called "algorithm for

Received November 18, 2014.
Accepted March 13, 2015.
[1]Corresponding author: fragomen@uga.edu

proven andyoung animals" (**APY**). This algorithm conditioned "young" animals on a small subset of "proven" animals. The APY algorithm has a cubic cost with the number of animals treated as proven and a linear cost with the animals treated as young; direct inversion is required for only a small portion of **G** composed of relationships among animals treated as proven. This algorithm was tested with simulated data and with US Holstein data (Fragomeni et al., 2014). In simulations, accuracies with APY were close to those with direct inverted **G** even when some animals with records were treated as young. This suggests that the definition of "proven" is not critical and this subset may not need to be composed of parents or animals with records, or possess any other special requirement. In US Holsteins with genotypes on 15k proven bulls and 60k young bulls, the correlations of genomic EBV (**GEBV**) obtained through APY and regular method were >0.99.

In real data sets, genotyped animals include bulls and cows. Although the number of proven bulls is limited and increases slowly (~2,000/yr for US Holsteins), the number of cows with genotypes can be very high. The purpose of this study was to evaluate the accuracy of GEBV with APY for US Holsteins, considering genotypes of bulls and cows and treating various groups of animals as proven and young.

MATERIALS AND METHODS

Genomic Recursions

The recursion for the additive genetic effect of animal i (u_i) can be written as (Misztal et al., 2014)

$$u_i \mid u_1 ... u_{i-1} = \sum_{j=1}^{i-1} p_{ij} u_j + \varepsilon_i,$$

where u is an additive genetic effect, p relates animals to all previous j individuals, and ε is the error term. Calculations can proceed as

$$\mathbf{p}_{i,1:i-1} = \mathbf{g}_{i,1:i-1} \left(\mathbf{G}_{1:i-1,1:i-1} \right)^{-1},$$
$$\mathbf{M}_{i,i} = m_i = \mathrm{var}\left(\varepsilon_i\right) = g_{i,i} - \mathbf{p}_{i,1:i-1} \mathbf{g}'_{i,1:i-1},$$

where **M** is a diagonal matrix of genomic Mendelian sampling, and $\mathbf{G} = \{g_{ij}\}$ is a genomic relationship matrix, **g** is a vector of **G**, and **p** is a vector that contains p. Then, the inverse of **G** can be created using a formula as in Henderson (1976) and Quaas (1988):

$$\mathbf{G}^{-1} = \left(\mathbf{I} - \mathbf{P}\right)' \mathbf{M}^{-1} \left(\mathbf{I} - \mathbf{P}\right) = \mathbf{T}' \mathbf{M}^{-1} \mathbf{T},$$

where **T** is a triangular matrix, $\mathbf{P} = \{p_{ij}\}$, and **I** is an identity matrix; if many of its elements are very small, they can be set to 0 and $\mathbf{G}^{-1}$ may be computed at a low cost.

The APY Algorithm

In genomic recursions, contributions from proven and young animals can be separated as

$$u_i \mid u_1, u_2, ..., u_{i-1} = \sum_{j \in "proven"} p_{ij} u_j + \sum_{j \in "young"} p_{ij} u_j + \varepsilon_i.$$

However, the contribution of information from young animals to other genotyped animals is 0 in GBLUP because young animals do not get information from data. Then, neglecting these contributions,

$$u_i \mid u_1, u_2, ..., u_{i-1} = \sum_{j \in "proven"} p_{ij} u_j + \varepsilon_i.$$

As shown in Misztal et al. (2014), the simplified recursions lead to a new formula for an approximate inverse of **G** called APY:

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

$$m_{g,i} = g_{ii} - \mathbf{G}_{ip} \mathbf{G}_{pp}^{-1} \mathbf{G}_{pi},$$

where $\mathbf{G}_{pp}$ is a subset of **G** relating proven animals, $\mathbf{G}_{py}$ relates proven and young animals, $\mathbf{G}_{ip}$ relates the ith young animal with all proven animals, and $\mathbf{M}_g$ is a diagonal matrix. Although this algorithm results in the same GEBV for GBLUP as the regular inversion of G^{-1}, for ssGBLUP, the APY algorithm leads to an approximation, as a young genotyped animal may provide ties to ungenotyped ancestors. This happens if at least one of its parents is not genotyped.

The APY $\mathbf{G}^{-1}$ is a sparse matrix with nonzero elements forming an L shape, with only a diagonal for the submatrix due to young animals; the only direct inversion required is for $\mathbf{G}_{pp}$. Whereas the regular $\mathbf{G}^{-1}$ requires quadratic storage and cubic computations, the APY $\mathbf{G}^{-1}$ requires quadratic storage and cubic computations only for animals treated as proven and linear storage and computations for animals treated as young. When the number of animals treated as proven is a small fraction of all animals, the APY $\mathbf{G}^{-1}$ has approximately a linear cost and can provide large savings in memory and especially in computing time.

Field Data

To check the quality of this approximation for $\mathbf{G}^{-1}$, we tested it using real data. Phenotypic data included 11,626,576 records for final score on 7,093,380 cows, with 10,709,878 animals in the pedigree provided by Holstein Association USA Inc. (Brattleboro, VT). Final score is a weighted linear combination of 5 major breakdown score for type traits in dairy cattle, and was chosen for this study because of availability of records. Genotypes on 42,503 SNP markers were available for 569,404 animals. However, to have comparisons with the regular ssGBLUP where direct inversion of $\mathbf{G}$ is used, analyses involved only 100,000 of the genotyped animals, which is the limitation of ssGBLUP for the available computer. Thus, genotypes were considered for all 23,174 bulls with progeny information, all 27,215 cows with records (hereafter termed "cows"), and additionally 49,611 young animals.

Analyses

Initially, GEBV were calculated using the regular ssGBLUP, which applies direct inversion for $\mathbf{G}$. Second, GEBV were calculated using APY to obtain $\mathbf{G}^{-1}$ recursively ($\mathbf{G}^{-1}{}_{\text{APY}}$) with several different definitions for proven animals: only sires; sires and cows; only cows; and sires with >5 progeny including sons and daughters. Third, previous analyses were repeated with "proven" animals randomly sampled from the group of all 100k genotyped animals in sets of 2k, 5k, 10k, 15k, and 20k animals; the sampling was replicated 4 times. Evaluations for final score were done using a single trait model as described in Tsuruta et al. (2002). All analyses were conducted with blup90iod2 (http://nce.ads.uga.edu/wiki/BLUPmanual) program with modifications as in Aguilar et al. (2011). The quality of approximations was assessed by correlations between GEBV for the almost 50k young animals obtained from ssGBLUP using direct inversion of full $\mathbf{G}$ (regular ssGBLUP) and ssGBLUP using approximated $\mathbf{G}^{-1}$ from the APY algorithm.

RESULTS AND DISCUSSION

Table 1 summarizes runs with regular and APY ssGBLUP when the subset of animals treated as "proven" were sires, sires + cows, cows, and sires with >5 daughters. For all subsets, the correlations of GEBV obtained with a regular and APY algorithms are >0.99. In all cases except when cows were treated as proven, the convergence rate was close to a regular run, indicating good computing properties. The smallest set of proven animals with good predictive ability was sires with >5 daughters (16,434 animals). Treating more animals as proven—that is, including sires with <5 progeny—only marginally affected the correlations. Computing an inverse for 16k animals (assuming cubic algorithm for inversion) cost about 200-fold less than for 100k animals and would cost 4,000-fold less for 600k animals.

Table 1. Correlations between genomic EBV with regular and APY (algorithm for proven and young) single-step genomic BLUP for young genotyped animals and rounds to convergence for different subset of animals used in recursions

Definition of subset	Animals in subset	Correlation	Rounds to convergence
All	100,000	1.000	567
Sires	23,174	0.994	432
Sires + cows	50,389	0.995	428
Cows	27,215	0.992	797
Sires >5 progeny	16,434	0.992	415

Surprisingly good correlations were observed with only cows treated as proven although the convergence rate was affected, but was still much better than with ssGBLUP with unsymmetric equations constructed to avoid the inverse of $\mathbf{G}$. (Aguilar et al., 2013). This means that the original definition of animals as young and proven is not necessarily important for accuracy of GEBV; only the number of animals in $\mathbf{G}_{pp}$ matters. To test this hypothesis, 2k, 5k, 10k, 15k, and 20k animals were chosen randomly from all bulls and cows and treated as proven in the APY algorithm. Rounds to convergence increased with subset size but were lower than with the regular algorithm. This suggest that $\mathbf{G}^{-1}$ by APY is well numerically conditioned. The correlations of GEBV with the regular and APY algorithms ranged from >0.94 for 2k animals to >0.99 for 20k animals, with very small variations among the replicates (Table 2). This means that the choice of animals in $\mathbf{G}_{pp}$ is mostly arbitrary.

Initially, the last statement seems hard to believe; however, recursions generate very similar inverses regardless of the order of animals. The single step modi-

Table 2. Ranges of correlations between genomic EBV with regular and APY (algorithm for proven and young) single-step genomic BLUP for young genotyped animals and rounds to convergence when different numbers of randomly sampled animals were used in the subset for recursions

Number of proven animals	Correlation	Rounds to convergence
2,000	0.943–0.944	351–357
5,000	0.971–0.972	354–367
10,000	0.985	391–403
15,000	0.989–0.990	411–480
20,000	0.992–0.993	416–425
20,000[1]	0.989–0.990	552–556

[1]Proven were randomly sampled from the group of young animals.

fies the pedigree relationship matrix ($\mathbf{A}$) toward a realized relationship matrix ($\mathbf{H}$). Possibly, to obtain a good $\mathbf{H}$, only a good sample of genotyped animals is needed, and several such samples may exist.

To test whether the presence of sires and cows is crucial for good properties of APY, an extra set included 20k animals selected randomly only from young animals. The correlations of GEBV for this set were slightly lower than with complete random 20k choice and similar to a 15k random sample. Also, the convergence rate was slightly worse. In general, we expect better properties of APY when animals treated as "proven" are well related to animals treated as "young." Although proven sires are well related to the general population, cows and young animals may be less so.

The Henderson's algorithm for creating the inverse of the numerator relationship matrix ($\mathbf{A}^{-1}$) is based on younger animals conditioned on older animals (Henderson, 1976). In such a case, each recursion has at most 2 nonzero elements, each with a value of 0.5 and due to a parent. However, an identical $\mathbf{A}^{-1}$ can be derived with animals in the reverse order (see Appendix in Misztal et al., 2014). In such a case, the number of nonzero elements in each recursion can be greater than 2 and they can take different values. Assume the following genomic recursion, where the additive genetic effect of an animal i is conditioned on the first m animals:

$$u_i|u_1,\ldots,u_{i-1} = \sum_{j=1}^{\min(i-1,m)} p_{ij}u_j + \varepsilon_i(\mathrm{m}),$$

where $\varepsilon_i(m)$ is the error term; although the error term should be smaller with larger m, apparently the reduction of $\varepsilon_i(m)$ for $m >$10k is small. In an alternate interpretation, the inverse of $\mathbf{G}$ created with APY is becoming more accurate as m increases, with small improvements beyond $m >$ 10k.

The limited number of animals required in the recursion (<20k) suggests that the genomic information for a population has a limited dimensionality (<20k). Nearly all genomic information from a reference population is usually assumed to be accounted for by SNP solutions with a medium-size chip (~50,000 SNP markers). However, many SNP are correlated. Pintus et al. (2013) found that 15,207 principal components extracted from matrices based on 39,555 SNP markers explained 99% of the genetic variation. Thus, the real dimensionality of the SNP information may be ~15k. Alternately, when the number of QTL is high, the accuracy of GBLUP depends on the number of independent chromosome segments, with the number of the segments usually <10k (Daetwyler et al., 2010). Further research will determine whether the limits based on the recursion, eigenvalues, and chromosome segments are related through equivalent models.

The US Holstein population is very homogeneous. In other species, populations may be more diverse and a larger subset may be needed. Lourenco et al. (2015) applied APY to genetic evaluation of US Angus for 3 traits with 52k genotyped animals. Using 4k and 8k subsets generated 84 and 97% of gains in accuracy, respectively, over BLUP compared with a regular ssGBLUP. A detailed analysis on the number and choices of animals treated as proven in APY will be a topic for a separate study. Further investigations will also look at whether specific subgroups of animals are invariant to the selection of the subset of animals defined as proven.

The original derivation of the APY algorithm was based on labeling animals in the recursion as proven. Because the algorithm works with any sufficiently large subset of animals in the recursion, the designation of proven or young may no longer be relevant. In particular, the animals can be decomposed into a base genomic relationship group (b) and a conditional genomic relationship group (c); for example, with relevant matrices $\mathrm{G_{bb}}$ and $\mathrm{G_{bc}}$.

In this paper, we focused on accuracy of ssGBLUP with $\mathbf{G}^{-1}$ calculated by APY. In practical implementations, important issues will be memory requirements and computing costs for a large number of genotyped animals. Assume a total of n = 500k genotyped animals, recursion on m = 20k animals, and double precision half-storage. The amount of memory necessary for APY $\mathbf{G}^{-1}$ is approximately 80 gigabytes (n × m × 8 = 20k × 500k × 8 bytes) or 8% of 1 terabyte ($\mathrm{n^2}/2 \times 8$ = 500k × 500k × 8 bytes/2) required for a regular half-stored $\mathbf{G}^{-1}$. As current servers have memory capacity in the order of terabytes, the memory requirements will not limit the APY algorithm. Computing APY $\mathbf{G}^{-1}$ would require approximately $\mathrm{m^3} + 2\mathrm{m^2(n - p)}$ operations, or about 0.3% operations ($\mathrm{n^3}$) for a regular $\mathbf{G}^{-1}$. Another issue is efficient computations of $\mathbf{A}_{22}^{-1}$. In separate analyses (results not provided), computing this matrix using formulas similar to that of Stranden and Mantysaari (2014) took negligible time and memory.

CONCLUSIONS

Inverse of a genomic relationship matrix can be approximated with the APY algorithm where actual inversion is applied only to a small subset of genotyped "proven" animals and an approximate inversion by recursion is applied on "young" animals. The approximation is very accurate when the number of animals in the subset is 10k or greater, and storage and computing costs can be dramatically lower. The

choice of animals in the subset is arbitrary, as various definitions, including random choices, provide similar accuracy. The convergence rate is superior to that of conventional inversion. Costs of APY inversions with a larger number of animals are approximately linear, making the algorithm potentially suitable for any number of genotypes. Single-step GBLUP with APY may be suitable for models with any number of genotyped animals.

ACKNOWLEDGMENTS

We gratefully acknowledge the very helpful comments by the two anonymous reviewers and insightful suggestions by section editor Jennie Pryce. This research was supported by grants from Zoetis (Florham Park, NJ), Cobb-Vantress Inc. (Siloam Springs, AR), Smithfield Premium Genetics (Rose Hill, NC), American Angus Association (St. Joseph, MO), Holstein Association USA (Brattleboro, VT), Pig Improvement Company (Hendersonville, TN), 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 (Washington, DC). A. L. thanks INRA metaprogram SelGen and projects X-Gen and GenSSeq.

REFERENCES

Aguilar, I., A. Legarra, S. Tsuruta, and I. Misztal. 2013. Genetic evaluation using unsymmetric single step genomic methodology with large number of genotypes. Interbull Bull. 47:222–225.

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.

Christensen, O. F., and M. S. Lund. 2010. Genomic predictions when some animals are not genotyped. Genet. Sel. Evol. 42:2.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031.

Faux, P., N. Gengler, and I. Misztal. 2012. A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix. J. Dairy Sci. 95:6093–6102.

Fernando, R. L., J. C. M. Dekkers, and D. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol. 46:50.

Fragomeni, B. O., I. Misztal, D. A. L. Lourenco, S. Tsuruta, Y. Masuda, and T. J. Lawlor. 2014. Use of genomic recursions and algorithm for proven and young animals for single-step genomic BLUP analyses with a large number of genotypes. Abstract 509 in Proc. 10th WCGALP, Vancouver, Canada. Access Mar. 31, 2015. https://asas.org/docs/default-source/wcgalp-posters/509_paper_9812_manuscript_1507_0.pdf?sfvrsn=2.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–93.

Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livest. Prod. Sci. 166:54–65.

Legarra, A., and V. Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in single-step best linear unbiased prediction. J. Dairy Sci. 95:4629–4645.

Liu, Z., M. E. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with direct estimation of marker effects. J. Dairy Sci. 97:5833–5850.

Lourenco, D. A. L., S. Tsuruta, B. Fragomeni, I. Aguilar, Y. Masuda, J. K. Bertrand, and I. Misztal. 2015. Genomic evaluation by single-step GBLUP in Angus. J. Anim. Sci. (Accepted).

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.

Pintus, M. A., E. L. Nicolazzi, J. B. C. H. M. Van Kaam, S. Biffani, A. Stella, G. Gaspa, C. Dimauro, and N. P. P. Macciotta. 2013. Use of different statistical models to predict direct genomic values for productive and functional traits in Italian Holsteins. J. Anim. Breed. Genet. 130:32–40.

Quaas, R. L. 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71:1338–1345.

Stranden, I., and E. A. Mantysaari. 2014. Comparison of some equivalent equations to solve single-step GBLUP. In Proc. 10th WCGALP, Vancouver, Canada. Accessed Mar. 31, 2015. https://asas.org/docs/default-source/wcgalp-proceedings-oral/069_paper_9344_manuscript_568_0.pdf?sfvrsn=2.

Tsuruta, S., I. Misztal, L. Klein, 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.

VanRaden, P. M., and J. R. Wright. 2013. Measuring genomic pre-selection bias 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.