Use of Genomic Recursions and Algorithm for Proven and Young Animals for Single-Step Genomic BLUP Analyses with a Large Number of Genotypes

***B.O. Fragomeni*[*]; I. Misztal[*]; D.A.L. Lourenco[*]; S. Tsuruta[*], Y. Masuda[*†], T. J. Lawlor[‡]**
[*]*University of Georgia, Athens, GA;* [†]Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Japan; [‡]Holstein Association USA Inc., Brattleboro, VT 05302

ABSTRACT: The purpose of this study was to evaluate 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 the inversion of G by genomic recursions. With efficient implementation, the algorithm has a cubic cost for proven animals but only a linear cost for young animals. Analyses involved simulated genomic data consisting of 20k genotyped animals for 45k SNP and real final score data for 74,980 genotyped Holstein bulls. The correlation between APY and regular genomic EBV of genotyped animals was >0.96 for simulated data under selection. For the real data, the correlation was >0.99. The APY algorithm may allow using all the available genotypes in one ssGBLUP analysis to reduce biases due to preselection of young animals.
Key words: single-step method; genomic selection

Introduction

Single-step genomic BLUP (Aguilar et al. (2010)) emerged as simple yet accurate tools for genetic evaluation. With implementation in the blupf90 suite (Misztal et al. (2002)), it supported a large 18-trait evaluation with 35k genotyped animals (Tsuruta et al. (2011)). As defined, the single-step needs inverses of the genomic (**G**) and pedigree ($\mathbf{A_{22}}$) relationship matrices for genotyped animals. With algorithms as in Aguilar et al. (2011), the cost of obtaining these matrices is cubic, and currently software has a limit of approximately 100k genotypes in the model. Several approaches were proposed to overcome such a limit, but they either had convergence problems or were expensive and hard to use for realistic models. Recently Misztal et al. (2014) proposed a method based on genomic recursion, where genomic EBV (GEBV) of a new genotyped animal is conditioned on GEBV of all the previous genotyped animals. One of their proposed algorithms for proven and young animals (APY) had cubic cost with the proven animals and linear cost with the young animals. The purpose of this study is to evaluate this algorithm with simulated and field data sets.

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)):

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

where

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

and **G** is a genomic relationship matrix. Then, the inverse of **G** can be created using a formula as in Henderson (1976) and Quaas (1988):

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

The APY algorithm. Partitions animals into proven and young. Then the recursion is:

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

In GBLUP, the contributions from young animals are 0 and:

$$u_i|u_1,u_2,\mathrm{K}\ ,u_{i-1} = \sum_{j\in\text{"}proven\text{"}} p_{ij}u_j + \varepsilon_i$$

Simplified recursions lead to the APY algorithm:

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

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

Assuming $\mathbf{G}=\mathbf{Z}\mathbf{Z}'/\mathrm{q}$, where **Z** is a matrix of genotypes and q is a normalizing constant, the APY algorithm can be expressed as:

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}_{pp}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{pp}^{-1}\mathbf{Z}_p\mathbf{Z}'_y/\mathrm{q} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_g^{-1} \begin{bmatrix} -\mathbf{Z}_y\mathbf{Z}'_p/\mathrm{q}\,\mathbf{G}_{pp}^{-1} & \mathbf{I} \end{bmatrix},$$

$$m_{g,i} = g_{ii} - \mathbf{z}'_i\mathbf{Z}'_p\mathbf{G}_{pp}^{-1}\mathbf{Z}_p\mathbf{z}_i/q^2$$

In the last formula, **Z** can be stored at low precision (even one byte integer), for lower memory requirements. When used in an iteration-on-data with the PCG

algorithm (Tsuruta et al. (2001)), costs with the APY algorithm are cubic for proven animals and linear for young animals. While the APY algorithm calculates the same GEBV for GBLUP (BLUP when the left hand side consists of G matrix), for ssGBLUP the GEBV calculated is an approximation.

Simulated data. The populations were simulated using QMSim (Sargolzaei and Schenkel (2009)). A single trait was simulated with heritability of 0.3. In each replicate 295k phenotypes were simulated, under a relationship structure of the same size over 20 generations. Also, 20k of these animals were genotyped for 45k SNP; the simulated genomic data mimicked the bovine genome. The simulation assumed BLUP selection with high and low intensity in a closed population.

Half of the genotyped animals were considered young and were selected from the last generation. In the first scenario, no records were used for these 10k young animals (which totalized 272k phenotypes in the data file). In the second scenario, all young animals had records (which totalized 295k phenotypes in the data file).

For the genotyped animals considered as proven, 10k animals were randomly sampled among those with at least 3 progenies.

Field data. Data for 10,102,702 Holsteins including 6,930,618 cows with records for final score (h^2=0.43). A total of 74,980 genotyped bulls were used in the ssGBLUP analysis to compare the breeding values estimated with regular and APY **G** inversion. Different approaches were used to classify the animals as young and proven: year of birth (2007 to 2009 and 2011) and presence of offspring (both male and/or female).

The data was also truncated to year ≥ 1985; a similar truncation did not reduce the accuracy for young animals while reducing computations (Lourenco et al. (2014)). With truncation, the number of animals in the relationship matrix decreased to 7,658,885, and the number of cows with records was 5,454,533.

Single step GBLUP analyses. For the simulated data, the analysis was performed with ssGBLUP in a model that includes the overall mean as a fixed effect and random additive genetic and residual effects. This analysis included the regular $\mathbf{G}^{-1}$ and the APY inverted **G** matrix ($\mathbf{G}^{-1}_{APY}$). For the simulated data, both (regular and APY) analyses were performed for all young animals with and without records. The Holstein analysis used the same model as in Tsuruta et al. (2002).

Results and Discussion

Correlation and accuracy for simulated data. Tables 1 and 2 show correlations between GEBV of regular $\mathbf{G}^{-1}$ and $\mathbf{G}^{-1}_{APY}$ analyses under high and low intensity of selection, respectively. Accuracies (correlation between true and estimated breeding value) for populations under high and low intensity of selection are presented in Tables 3 and 4, respectively. While the correlations were high but < 1.0, the accuracies were similar for the two methods. Initially, the accuracies with the APY algorithm were lower with BLUP selection, but became equal when **G** was scaled to match $\mathbf{A}_{22}$ (Option tunedG 4 in the preGSf90 program; see notes by I. Aguilar at http://nce.ads.uga.edu/wiki/doku.php). This scaling of **G** in ssGBLUP was found effective under strong selection in simulated data by Vitezica et al. (2011). Yet, it was less important in chicken data, probably due to weak selection and a different population structure (Chen et al. (2011)).

Table 1. Correlation between regular and APY[1] genomic estimated breeding values for all, young, and proven animals in simulated data under high intensity of selection, for young animals with and without observations (obs.).

	G^2 without scaling		G with scaling	
Animals	No obs.	10k obs.	No obs.	10k obs.
All	0.90	0.94	0.99	0.99
Proven	0.99	0.99	1.00	1.00
Young	0.72	0.79	0.96	0.96

[1]APY – Algorithm for proven and young animals
[2]G – genomic relationship matrix

Table 2. Correlation between regular and APY[1] genomic estimated breeding values for all, young, and proven animals in simulated data under low intensity of selection, for young animals with and without observations (obs.).

	G^2 without scaling		G with scaling	
Animals	No obs.	10k obs.	No obs.	10k obs.
All	0.98	0.98	1.00	1.00
Proven	1.00	1.00	1.00	1.00
Young	0.91	0.94	0.99	0.98

[1]APY – Algorithm for proven and young animals
[2]G – genomic relationship matrix

Table 3. Accuracy (correlation between estimated and true breeding value) for regular and APY[1] genomic estimated breeding values for all, young, and proven animals in simulated data under high intensity of selection, for young animals with and without observations.

		No observations		10k observations	
G^2	Animals	Regular	APY	Regular	APY
Without scaling	All	0.69	0.55	0.85	0.79
	Proven	0.77	0.78	0.83	0.83

	Young	0.19	0.10	0.58	0.47
With scaling	All	0.83	0.83	0.86	0.86
	Proven	0.83	0.83	0.84	0.84
	Young	0.45	0.43	0.62	0.60

[1]APY – Algorithm for proven and young animals
[2]G – genomic relationship matrix

Table 4. Accuracy (correlation between estimated and true breeding value) for regular and APY[1] genomic estimated breeding values for all, young, and proven animals in simulated data under low intensity of selection, for young animals with and without observations.

		No observations		10k observations	
G^2	Animals	Regular	APY	Regular	APY
Without scaling	All	0.78	0.74	0.85	0.83
	Proven	0.81	0.81	0.84	0.83
	Young	0.45	0.39	0.68	0.64
With scaling	All	0.83	0.83	0.86	0.86
	Proven	0.83	0.83	0.85	0.84
	Young	0.61	0.60	0.73	0.72

[1]APY – Algorithm for proven and young animals
[2]G – genomic relationship matrix

Correlation for real data. The correlation between GEBV with $\mathbf{G}^{-1}_{APY}$ and the regular $\mathbf{G}_{-1}$ was >0.99 in all the cases (Table 5). The convergence was slow when the $\mathbf{G}^{-1}_{APY}$ was used. This problem occurred solely in the real data. When only animals with progeny were considered as proven, the accuracy was still high, and the convergence rate was similar to the regular analysis—574 rounds in the regular $\mathbf{G}^{-1}$ and 556 rounds in $\mathbf{G}^{-1}_{APY}$. Cutting the data before 1985 reduced the number of iterations to 489 in the regular $\mathbf{G}^{-1}$ and 487 in $\mathbf{G}^{-1}_{APY}$—and the correlation remained >0.99.

Table 5. Correlation between the regular and APY[1] genomic estimated breeding values, number of young animals and number of rounds to single step genomic BLUP convergence for the Holstein dataset, for different year definitions for proven animals or animals with progeny treated as proven.

Year	# young	All	Young	Proven	# rounds
2011	2,260	0.998	0.995	0.998	672
2009	45,428	0.999	0.999	0.999	963
2008	52,258	0.999	0.998	0.999	864
2007	56,872	0.998	0.997	0.999	847
Progeny	60,375	0.998	0.997	0.999	556

[1]APY – Algorithm for proven and young animals

Additional issues. In this study, we used a regular algorithm (direct inversion) for $\mathbf{A}_{22}^{-1}$, which is not applicable for a large number of genotypes. When we attempted to use the APY-equivalent algorithm for $\mathbf{A}_{22}^{-1}$, accuracies dropped. In particular, for the Holstein data set, the correlations dropped to 0.978 for proven animals and 0.843 for young animals, with convergence in 575 rounds after cutting data. Further studies will look into alternate ways of obtaining $\mathbf{A}_{22}^{-1}$ at a low cost, such as discussed in Faux and Gengler (2013) or Misztal et al. (2014).

Conclusion

The algorithm for proven and young animals provides reliable estimates for breeding values with reasonable computation cost. In order to achieve better convergence rate, proven and young animals should be classified correctly.

Literature Cited

Aguilar, I., Misztal, I., Johnson, D.L. et al. (2010). *J. Dairy. Sci.* 93:2, 743-752.

Aguilar, I., Misztal, I., Legarra, A. et al. (2011). *J. Anim. Breed. Genet.* 128: 422–428.

Chen, C.Y., Misztal, I., Aguilar, I. et al. (2011). *J. Anim. Sci.* 89:2673-2679.

Henderson, C. R. (1976). Biometrics. 32:6

Lourenco, D.A.L., Misztal, I., Tsuruta, S. et al. (2014) *J. Dairy. Sci.* (Accepted)

Misztal, I., Legarra, A., and Aguilar I. (2014). *J. Dairy. Sci.* (Accepted)

Faux, P. and Gengler, N. (2013). *Genet. Sel. Evol.* 45:45.

Quaas, R. L. (1988). J. Dairy Sci. 71:1338-1345.

Tsuruta S. and Misztal I. (2008). *J. Anim. Sci.* 86:1514-1518.

Sargolzaei, M. and Schenkel. F.S. (2009). *Bioinformatics*, 25:680-681.

Tsuruta, S., I. Misztal, Klei, L. et al. (2002). *J. Dairy. Sci.* 85:1324-1330.

Tsuruta, S., Misztal, I., and Aguilar, I. (2011). *J. Dairy. Sci.* 94:4198-4204.

Tsuruta, S., Misztal, I., and Stranden, I. (2001). *J. Anim. Sci.* 79:1166-1172.

Vitezica, Z.G., Aguilar, I., Misztal, I. et al. (2011). *Genet. Res.* 93:357-366.