Efficient Inversion of Genomic Relationship Matrix by the Algorithm for Proven and Young (APY)

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

[1] *Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA*
[2] *Instituto Nacional de Investigacion Agropecuaria, Canelones, 90200, Uruguay*
[3] *INRA, UMR1388 GenePhySE, Castanet Tolosan, 31326, France*
[4] *Holstein Association USA Inc., Brattleboro, VT 05302, USA*

Abstract

The purpose of this study was to evaluate properties of the inverse of the genomic relationship matrix derived with the algorithm for proven and young (APY) and the accuracy of genomic selection in single-step genomic best linear unbiased prediction (ssGBLUP). The APY implements genomic recursions on a subset of genotyped animals. When that subset is small, the cost of APY is approximately linear in memory and computations, effectively removing restrictions on the number of genotypes. Tests involved 10 102 702 final scores from 6 930 618 Holstein cows. A total of 100 000 animals with genotypes were used in the analyses and included 23 174 sires, 27 215 cows and 49 611 young animals. Genomic estimated breeding values (GEBVs) were calculated using ssGBLUP with a regular inverse of the genomic relationship matrix (**G**) and with **G** inverse from APY. Many subsets were tested including only sires, only cows and random samples from 2 000 to 20 000 animals. When the number of animals in the subset was ≥15,000, the correlations between GEBV with APY and GEBV with the regular inverse were ≥0.99. Best convergence rate was achieved with random samples. A theory on APY was derived and is based on the fact that additive effects of animals in the subset are linear functions of the effects of independent chromosome segments (ICSs); the number of segments is a function of the effective population size. Accuracy of GEBV with APY can be slightly superior to that of a regular inverse. The inverse with APY is computed from **G**, which in turn is derived from single nucleotide polymorphism (SNP) BLUP and indirectly from BayesB or other SNP-based prediction methods. Strategies like SNP selection, SNP weighting, and use of causative SNPs from sequence analysis can be incorporated in APY without additional cost. The APY removes size limitations from ssGBLUP and facilitates a model with a complex genetic architecture.

Key words: big population, genomic recursion, inversion, single-step method

Introduction

The ssGBLUP method (Aguilar *et al.*, 2010; Christensen and Lund, 2010) is an attractive tool for genetic evaluation. If the current evaluation is based on traditional BLUP, all that is needed to move from regular to genomic evaluation is to change a relationship matrix. However, with the current implementation of ssGBLUP, the number of genotyped animals was limited by costs to invert **G**. The current limit is about 100 000, but the U.S. Holstein industry has collected genotypes for almost a million animals.

Past progress in animal breeding was greatly due to a fast algorithm to invert the numerator relationship matrix (**A**; Henderson, 1976). Although the cost of inverting **A** is cubic with the number of animals, the cost of inversion using Henderson's algorithm is very low because the recursion includes at most two terms for an animal (one for its sire and one for its dam). More complicated recursions (eight terms) allowed for efficient computing of dominance relationships (Hoeschele and VanRaden, 1991).

When recursion is based on a limited number of individuals, the cost of inverting **G** can be lower. Misztal *et al.* (2014) postulated recursions on proven animals (with phenotypes or progeny) and called the methodology an algorithm for proven and young animals

(APY). This algorithm was tested in a population of 100 000 genotyped Holsteins and different groups of animals in recursions (Fragomeni *et al.*, 2015). They found that recursions on about 10 000 animals resulted in similar accuracy for GEBVs as with a regular inverse and that the choice of animals in recursions was unimportant.

The computing and storage costs are almost linear in APY, which allow inverting **G** of practically any size. However, why APY works and whether it has possible internal limitations have not been addressed. The purposes of this paper are to 1) present the formulas for APY, 2) present partial theory about why APY works, 3) present results of Fragomeni *et al.* (2015), and 4) demonstrate that APY is useful with SNP selection and causative SNP identified.

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)

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

where *p* relates animals to all previous individuals and ε is the error term. If **G** is available,

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

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

$$\mathbf{G}^{-1} = (\mathbf{I} - \mathbf{P})'\mathbf{M}^{-1}(\mathbf{I} - \mathbf{P}),$$

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

APY Algorithm

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

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

However, the contribution of information from young animals to other genotyped animals is 0 in GBLUP. Then, neglecting these contributions,

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

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

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}_{bb}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{bb}^{-1}\mathbf{G}_{bc} \\ \mathbf{I} \end{bmatrix} \mathbf{M}^{-1} \begin{bmatrix} -\mathbf{G}_{cb}\mathbf{G}_{bb}^{-1} & \mathbf{I} \end{bmatrix};$$
$$m_i = g_{ii} - \mathbf{G}_{ib}\mathbf{G}_{bb}^{-1}\mathbf{G}_{bi}.$$

where *b* relates to proven animals, *c* relates to young animals, $\mathbf{G}_{bb}$ is a subset of **G** relating proven animals, $\mathbf{G}_{bc}$ relates proven and young animals and $\mathbf{G}_{ib}$ relates young animal *i* with all proven animals. When the number of animals treated as proven is a small fraction of all animals, APY has approximately a linear cost and can provide large savings in memory and especially in computing time.

Field Data and Analyses

Phenotypic data included 11 626 576 phenotypes for final score of 7 093 380 Holstein cows, with 10 709 878 animals in the pedigree. Comparisons included 100 000 animals (23 174 bulls, 27 215 cows and 49 611 young) genotyped for 42 503 SNP markers. Initial GEBVs were calculated using regular ssGBLUP with direct inversion of **G**. Then, GEBVs were calculated using APY inverse ($\mathbf{G}_{\text{APY}}^{-1}$) with four different definitions for proven animals: 1) only sires, 2) sires and cows, 3) only cows and 4) sires with >5 progeny (including sons and daughters).

Finally, previous analyses were repeated with proven animals randomly sampled from the group of all 100 000 genotyped animals in sets of 2 000, 5 000, 10 000, 15 000, and 20 000 animals; sampling was replicated four times. Evaluations for final score were computed using a single-trait model as described in Tsuruta *et al.* (2002). All analyses were done using blup90iod2 (http://nce.ads.uga.edu/wiki/BLUPmanual) with modifications as in Aguilar *et al.* (2011). Accuracy of APY was assessed by correlations between GEBVs for the almost 50 000 young animals obtained from ssGBLUP using either direct inversion of full **G** or $\mathbf{G}^{-1}_{\text{APY}}$.

Results and Discussion

Field Data and ssGBLUP

Table 1 summarizes computations with regular and APY ssGBLUP for the four subsets of proven animals. For all subsets, correlations of GEBVs from regular and APY algorithms were >0.99. In all cases except when cows were treated as proven, the convergence rate was close to regular computation, indicating good computing properties. The smallest set of proven animals with good predictive ability was sires with >5 progeny (16 434 animals). Treating more animals as proven (i.e., including sires with ≤5 progeny) only marginally affected correlations. Computing an inverse by APY for 16 000 animals (assuming a cubic algorithm for regular inversion) costs about 18-fold less for 100 000 animals and would cost 700-fold less for 600 000 animals.

Surprisingly good correlations were observed with only cows treated as proven. Although convergence rate was affected, it was still much better than with ssGBLUP with unsymmetric equations constructed to avoid the inverse of **G** (Aguilar *et al.*, 2013). This means that the original definition of animals as young and proven is not necessarily important for accuracy of GEBVs, and only the number of animals treated as "proven" matters. To tests this hypothesis, 2 000, 5 000, 10 000, 15 000 and 20 000 animals were randomly chosen from all bulls and cows and treated as proven in the APY algorithm.

Rounds to convergence increased with the subset size but were lower than with the regular algorithm. This suggests that $\mathbf{G}^{-1}_{\text{APY}}$ is well conditioned numerically. Correlations of GEBVs from regular and APY algorithms ranged from >0.94 for 2 000 animals to >0.99 for 20 000 animals, with very small variations among replicates (Table 2). This means that the choice of proven animals is mostly arbitrary.

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 core (*b*) animals in the subset and the remaining non-core animals (*c*).

Table 1. Correlations between GEBVs from regular and APY ssGBLUP for young genotyped animals and rounds to convergence by subset of animals used in recursions.

Subset	Animals in subset	Correlation	Rounds
All	100 000	1.000	567
Sires	23 174	0.994	432
Sires and cows	50 389	0.995	428
Cows	27 215	0.992	797
Sires with >5 progeny	16 434	0.992	415

Table 2. Ranges of correlations between GEBVs from regular and APY ssGBLUP for young genotyped animals and rounds to convergence by number of randomly sampled animals (*N*) used in the subset for recursions.

N	Correlation	Rounds
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[a]	0.989–0.990	552–556

[a]Randomly sampled from young animals.

Theory of APY

The limited number of animals required in the recursion (<20 000) suggests that the genomic information for a population has a limited dimensionality (<20 000). Stam (1980) proposed that in populations with limited effective population size, the number of ICSs (or M_e) is limited to $4N_eL$, where N_e is effective population size and L is genome length in Morgans. Other formulas for M_e have been suggested (e.g., Daetwyler *et al.*, 2010).

Let $\mathbf{s}$ be a vector of effects of n ICSs. Assume that these effects explain nearly all the additive variance. Let t_{ij} be a fraction of segment j in individual i, and assume that the value of t_{ij} is $t_{ij}s_j$. Then, $\mathbf{u} = \mathbf{T}\mathbf{s} + \boldsymbol{\varepsilon}$, where $\mathbf{T}$ is a matrix that relates $\mathbf{u}$ to chromosome segments and $\boldsymbol{\varepsilon}$ is the fraction of breeding value unexplained by SNP effects. Applications to farm animals using medium size SNP chips usually assume $\boldsymbol{\varepsilon} \approx \mathbf{0}$ (VanRaden, 2008; Goddard *et al.*, 2011). Divide individuals arbitrarily into two groups: core (b) and non-core (c):

$$\mathbf{u}_b = \mathbf{T}_b\mathbf{s} + \boldsymbol{\varepsilon}_b;$$
$$\mathbf{u}_c = \mathbf{T}_c\mathbf{s} + \boldsymbol{\varepsilon}_c.$$

Assume that the number of core animals is equal to M_e, $\mathbf{T}$ is full rank (no clones), and SNP effects nearly fully explain breeding value ($\boldsymbol{\varepsilon}_b \approx \mathbf{0}$) Then, $\mathbf{s} \approx \mathbf{T}_b^{-1}\mathbf{u}_b$, or the ICS information is practically equivalent to that in additive effects of M_e core animals. Substituting

$$\mathbf{u}_c = \mathbf{T}_c\mathbf{s} + \boldsymbol{\varepsilon}_c \approx \mathbf{T}_c\mathbf{T}_b^{-1}\mathbf{u}_b + \boldsymbol{\varepsilon}_c \text{ and}$$
$$\mathbf{u}_c = \mathbf{P}\mathbf{u}_b + \boldsymbol{\varepsilon}_c,$$

we obtain the recursion formula used to derive APY.

The theory can be extended to more or fewer core animals. Whereas the APY inverse with fewer core animals would result in less accurate GEBV, using more than the optimal number of core animals would only increase computations without affecting the accuracy. Assuming for Holsteins an N_e of 100 and an L of 30, the number of ICSs is 12 000 based on Stam's formula. This is very close to the minimum number of core animals needed to achieve correlations of ≥0.99 in this study.

Genetic Architecture and G

Although the derivations for APY included effects of ICSs, those effects are absent from the final APY formula, which depends on $\mathbf{G}$ only. Therefore, any information on the specific genetic architecture of a trait, if present, is included through $\mathbf{G}$. In the GBLUP case (assuming the same variance for all SNP loci), $\mathbf{G}$ can be derived from SNP BLUP as $\mathbf{ZZ}'/q$, where q is a scaling factor (VanRaden, 2008). For weighted SNP BLUP with $\text{var}(\mathbf{a}) = \mathbf{D}\sigma_a^2$, where $\mathbf{a}$ is a vector of breeding values, $\mathbf{D}$ is a diagonal matrix of weights and σ_a^2 is total genetic variance, $\mathbf{G}$ becomes $\mathbf{ZDZ}'/q$.

If SNP BLUP includes causative and other SNPs with known $\mathbf{D}$, an equivalent $\mathbf{G}$ can be constructed and subsequently an equivalent APY inverse. Particularly, if all (say n) causative SNPs are identified, the equivalent APY inverse would require recursion on n animals.

Is APY Inverse an Approximation?

In some of our simulations, use of the APY inverse resulted in slightly higher accuracy of GEBV (results not shown). This could be the result of multiple factors. First, if the theory for ICSs is correct, doing more computations than necessary only introduces numerical errors. Second, as APY inverse does not require a block of $\mathbf{G}$ from non-core animals (except diagonals), sampling error in that part of $\mathbf{G}$ (due to a finite number of SNPs) is propagated to a regular but not to an APY inverse.

APY and Admixed Populations

If each breed has different ICSs, an admixed population would include ICSs from every breed, and recursions in APY need to include enough individuals to account for ICSs of all the breeds. The recursions can be constructed

to ignore non-existent relationships by using only relevant individuals in recursions. For example, assume that a population contains three breeds (A, B and C) and all two-way crosses. Core individuals would be purebreds only, and breeding values of crossbreds would be linear combinations of ICSs of purebreds. Recursions for individuals of breed A would not contain any individuals from breed B or C. Similarly, recursions for cross A × B would contain only individuals from breeds A and B.

Conclusions

The inverse of **G** can be computed with APY by recursion on a subset of animals that is equal to the number of ICSs (about 10 000 for Holsteins). For large genotyped populations, the algorithm has approximately a linear cost and, therefore, is applicable to any population size. The inverse with APY can be more accurate than a regular inverse, and a specific SNP architecture can be considered. For admixed populations, a selective use of recursions can minimize nonexistent covariances across subpopulations.

Acknowledgements

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., Legarra, A., Tsuruta, S. & Misztal, I. 2013. Genetic evaluation using unsymmetric single step genomic methodology with large number of genotypes. *Interbull Bulletin 47*, 222–225.

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.

Aguilar, I., Misztal, I., Legarra, A. & Tsuruta, S. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *Journal of Animal Breeding and Genetics 128*, 422–428.

Christensen, O.F. & Lund, M.S. 2010. Genomic predictions when some animals are not genotyped. *Genetics Selection Evolution 42*, 2.

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

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 best linear unbiased predictor (BLUP) with a large number of genotypes. *Journal of Dairy Science 98*, 4090–4094.

Goddard, M.E., Hayes, B.J. & Meuwissen, T.H.E. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics 128*, 409–421.

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.

Hoeschele, I. & VanRaden, P.M. 1991. Rapid inversion of dominance relationship matrices for noninbred populations by including sire by dam subclass effects. *Journal of Dairy Science 74*, 557–569.

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.

Quaas, R.L. 1988. Additive genetic model with groups and relationships. *Journal of Dairy Science 71,* 91–98.

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical Research 35,* 131–155.

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

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