# IMPACT OF AN APPROXIMATE INVERSE OF THE GENOMIC RELATIONSHIP MATRIX FOR SINGLE-STEP EVALUATION OF AUSTRALIAN MEAT SHEEP

**K. Meyer and A.A. Swan**

Animal Genetics & Breeding Unit*, University of New England, Armidale, NSW, 2351 Australia

## SUMMARY

Common implementations of single-step genomic evaluation require the inverse of the genomic relationship matrix. Obtaining the inverse can become computationally prohibitive as its size increases. Stimulated by rapidly increasing numbers of genotyped animals, several procedures to approximate this inverse have been proposed. We examine the impact of two methods of approximation on predicted breeding values for a multi-breed population of Australian sheep. Results show that very high correlations with predictions using the full inverse can be achieved whilst reducing computational requirements. However, current levels of genotyping in our data were relatively low and results need to be validated as larger number of genotypes become available.

## INTRODUCTION

The single-step procedure for joint genetic evaluation of genotyped and non-genotyped animals (ssGBLUP) has become routine in many livestock improvement schemes. In essence, it extends the classic breeding value model to include genomic information by replacing the pedigree based relationship matrix ($\mathbf{A}$) with its counterpart ($\mathbf{H}$) which combines both. Only $\mathbf{H}^{-1}$ is required in the mixed model equations (MME) to be solved. This can be formed directly, but does require the inverse of two matrices of size $n_2 \times n_2$, with $n_2$ the number of genotyped animals. The first is the inverse of the dense genomic relationship, $\mathbf{G}$, which needs to be inverted explicitly. The second is the inverse of $\mathbf{A}_{22}$, the corresponding part of $\mathbf{A}$, which can be obtained indirectly by exploiting partitioned matrix results (e.g. Strandén *et al.* 2017). Rapidly increasing numbers of genotyped animals have stimulated development of approximations for $\mathbf{G}^{-1}$. We examine the impact of two proposed schemes for a multi-breed set of sheep data, namely the 'algorithm for proven and young' (APY) sires (e.g. Misztal *et al.* 2014) and the use of the Woodbury matrix identity combined with a reduction in the number of principal components (PCs) considered, dubbed TBLUP (Mäntysaari *et al.* 2017).

## MATERIAL AND METHODS

**The APY inverse.** Reorder and split $\mathbf{G}$ into a set of 'core' (or proven) animals and a set of 'non-core' (or young) animals, denoted by subscripts 'C' and 'N', respectively. This gives

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}_{CC}^{-1} + \mathbf{G}_{CC}^{-1}\mathbf{G}_{CN}\mathbf{G}^{NN}\mathbf{G}_{NC}\mathbf{G}_{CC}^{-1} & -\mathbf{G}_{CC}^{-1}\mathbf{G}_{CN}\mathbf{G}^{NN} \\ -\mathbf{G}^{NN}\mathbf{G}_{NC}\mathbf{G}_{CC}^{-1}\mathbf{G}_{NC} & \mathbf{G}^{NN} \end{bmatrix} \quad \text{for} \quad \mathbf{G} = \begin{bmatrix} \mathbf{G}_{CC} & \mathbf{G}_{CN} \\ \mathbf{G}_{NC} & \mathbf{G}_{NN} \end{bmatrix}$$

with $\mathbf{G}^{NN} = \left(\mathbf{G}_{NN} - \mathbf{G}_{NC}\mathbf{G}_{CC}^{-1}\mathbf{G}_{CN}\right)^{-1} = \mathbf{G}_{NN.C}^{-1}$, where $\mathbf{G}_{NN.C}$ is the matrix of relationships amongst non-core animals conditional on the core animals. For pedigree relationships, the diagonals of the corresponding function of $\mathbf{A}$ represent Mendelian sampling terms. Moreover, if non-core animals had no progeny, the matrix would be diagonal. Analogously, if non-core animals can be chosen so that $\mathbf{G}_{NN.C}$ is close to diagonal, a suitable approximation of $\mathbf{G}^{-1}$ can be obtained by substituting $\mathbf{D}_N = \text{Diag}\{\mathbf{G}_{NN.C}\}$ for it (Misztal *et al.* 2014). This gives an approximate inverse which is considerably sparser than $\mathbf{G}^{-1}$ and can reduce computational demands dramatically.

**The TBLUP inverse.** Consider $\mathbf{G}$ of form $(\lambda/s)\mathbf{Z}\mathbf{Z}' + \mathbf{B}$ with $\mathbf{Z}$ the $n_2 \times m$ matrix of $m$ centered marker counts and $s$ a scale factor. A common choice for $\mathbf{B}$ is $(1-\lambda)\mathbf{A}_{22} + \lambda\alpha\mathbf{J}$ for $\lambda < 1$, $\alpha \geq 0$ a

* A joint venture of NSW Department of Primary Industries and the University of New England

small constant and **J** a matrix with all elements equal to unity. The Woodbury identity gives

$$\mathbf{G}^{-1} = \mathbf{B}^{-1} - (\lambda/s)\mathbf{B}^{-1}\mathbf{Z}\Big(\mathbf{I} + (\lambda/s)\mathbf{Z}'\mathbf{B}^{-1}\mathbf{Z}\Big)^{-1}\mathbf{Z}'\mathbf{B}^{-1} = \mathbf{B}^{-1} - \mathbf{T}'\mathbf{T} \qquad \text{with } \mathbf{T} \text{ of size } m \times n_2.$$

Similarly, $\mathbf{B}^{-1} = (1-\lambda)^{-1}[\mathbf{A}_{22}^{-1} - \psi\mathbf{A}_{22}^{-1}\mathbf{J}\mathbf{A}_{22}^{-1}]$ with $\psi = \lambda\alpha/(1 - \lambda + \lambda\alpha\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1})$. This can reduce computational requirements to obtain $\mathbf{G}^{-1}$ if $m$ is substantially smaller than $n_2$. Further, let $(\lambda/s)\mathbf{Z}'\mathbf{B}^{-1}\mathbf{Z} = \mathbf{V}\mathbf{E}\mathbf{V}'$, where **E** denotes the diagonal matrix of eigenvalues and **V** the corresponding matrix of eigenvectors. An approximate inverse of **G** can then be obtained by considering the $r < m$ largest eigenvalues and corresponding eigenvectors only, i.e replacing **T** above with $\mathbf{T}_r = (\mathbf{E}_r + \mathbf{I}_r)^{-1/2}\mathbf{V}_r'\mathbf{Z}$, of size $r \times n_2$ (Mäntysaari *et al.* 2017).

**Data and model.** Data consisted of 1,206,908 records for eye muscle depth, recorded for Australian terminal sire sheep breeds between 1990 and 2018. These included Poll Dorset, Suffolk, White Suffolk and Texel as the main breed groups and 18 other, less numerous breeds. Breed differences were modeled by appropriately defined genetic group effects.

Data were pre-corrected for fixed effects of birth and rearing type, age, age of dam and body weight. The model of analysis comprised additive genetic effects (random) for 1,698,838 animals in the pedigree, 54,094 contemporary groups (fixed), 93 genetic groups (random) and 56,212 sire × flock-year (random) effects. Genotype information, comprised of marker counts for $m = 48{,}599$ SNPs, was available for 23,040 animals.

**Analyses.** The 'raw' genomic relationship matrix, was built using Method 1 of Van Raden (2008), $\mathbf{G}_M = \mathbf{Z}\mathbf{Z}'/s$, centering marker counts by observed gene frequencies. **G** was then formed as the weighted average of $\mathbf{G}_M$ and $\mathbf{A}_{22}$ aligning the matrices as described by Vitezica *et al.* (2011), $\mathbf{G} = \lambda(\mathbf{G}_M + \alpha\mathbf{J}) + (1-\lambda)\mathbf{A}_{22}$ for $\alpha = 0.02497$, and arbitrarily chosen weighting factor of $\lambda = 0.95$.

Analyses considered APY core sizes from $n_C = 2.5$K to 20K (with K denoting a factor of 1000). Core animals were chosen either by picking genotyped animals at random (RND) or by selecting those with the most progeny (PRG). TBLUP type approximations of $\mathbf{G}^{-1}$ utilised the leading PCs explaining between 90% and 99% of total variation. Single-step BLUP analyses were carried for all approximations of $\mathbf{G}^{-1}$ and contrasted to a 'standard' ssGBLUP analysis with the 'full' $\mathbf{G}^{-1}$ (FULL). MME were solved iteratively using a preconditioned conjugate gradient (PCG) algorithm with simple, diagonal preconditioner. All calculations were carried out using WOMBAT (Meyer 2007).

Summary statistics calculated were correlations between predicted total breeding values (EBV), i.e. the sum of the predicted additive genetic effects and the appropriate portions of the predicted genetic group effects, from FULL and APY or TBLUP analyses. In addition, corresponding regression coefficients and ranges of differences in EBVs were examined.

## RESULTS

Correlations between and regressions of EBVs from FULL on APY analyses are summarised in Table 1. As in various literature reports, there were only small differences between schemes to select core animals. Core sizes about 15K were required to ensure correlations for non-core animals to be close to 0.999. This is in line with results of Pocrnic *et al.* (2016a,b) who demonstrated for a number of livestock species that core sizes of 15K or less sufficed to achieve peak predictive accuracies. Based on simulations linking core and effective population size, the authors recommended a core size equal to the number of eigenvalues (of **G**) explaining 98% of total variation. For $\mathbf{G}_M$ and **G** this was equal to 15,220 and 16,714, respectively. In comparison, for a multi-breed population of New Zealand sheep, 18.8K eigenvalues were needed to capture 98% of the variation among 47K genotypes (Nilforooshan and Lee 2019). Linear regressions of FULL on APY EBVs for core sizes of 10K or more were essentially unity (with corresponding intercepts close to zero) demonstrating that approximation of $\mathbf{G}^{-1}$ at sufficient core size did not distort distributions of EBVs markedly.

**Table 1. Relationship between total predicted breeding values from single-step analyses using the 'full' inverse of the genomic relationship matrix and its APY approximation**

| Type[a] | Sel.[b] | Correlation | | | | Regression coefficient | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 2.5[c] | 5 | 10 | 15 | 2.5 | 5 | 10 | 15 |
| NOG | RND | 0.9993 | 0.9997 | 0.9999 | 1.0000 | 1.0021 | 1.0013 | 1.0009 | 1.0000 |
| | PRG | 0.9992 | 0.9997 | 0.9999 | 1.0000 | 0.9980 | 0.9999 | 1.0006 | 1.0002 |
| NOC | RND | 0.9644 | 0.9831 | 0.9953 | 0.9986 | 0.9974 | 1.0038 | 1.0040 | 1.0007 |
| | PRG | 0.9636 | 0.9833 | 0.9953 | 0.9988 | 0.9789 | 1.0063 | 1.0074 | 1.0048 |
| COR | RND | 0.9941 | 0.9983 | 0.9996 | 0.9999 | 0.9849 | 0.9921 | 1.0006 | 1.0003 |
| | PRG | 0.9991 | 0.9991 | 0.9997 | 0.9999 | 0.9854 | 0.9956 | 0.9985 | 1.0004 |

[a] NOG: non-genotyped, NOC: non-core and COR: core animals [b] Selection of core animals: RND random, PRG most progeny [c] Number of core animals; in thousand

Table 2 shows the numbers of non-zero elements in $\mathbf{H}^{-1}$ for different APY approximation of $\mathbf{G}^{-1}$ and their effects on the number of iterates required to solve the MME. In comparison, corresponding numbers for FULL, were 271 million elements and 611 iterates. Use of APY tended to increase the number of iterates required somewhat, especially when selecting core animals with most progeny. A similar increase over the standard ssGBLUP has been reported by others (Strandén *et al.* 2017; Mäntysaari *et al.* 2017).

For $n_2$ = 23,040 genotyped animals and $m$ = 48,599 SNPs considered, there was no computational advantage for the Woodbury inverse of $\mathbf{G}$. Moreover, the number of non-zero eigenvalues of $(\lambda/s)\mathbf{Z}'\mathbf{B}^{-1}\mathbf{Z}$ was limited to $n_2$. As shown in Table 3, sufficient PCs – just over 15K – to explain about 97% of total variation were required to yield correlations between TBLUP and FULL EBVs for genotyped animals of 0.999. Corresponding regression coefficients (not shown) were again close to unity. As for APY, there was a slight trend for the number of iterates to increase with less approximation, i.e. more PCs considered.

## DISCUSSION

Approximation of $\mathbf{G}^{-1}$ via APY is widely used and has made ssGBLUP for very large numbers of genotypes feasible. For instance, Lourenco *et al.* (2018) described the APY implementation for American Angus cattle with 450K genotyped animals, and Masuda *et al.* (2017) reported on dairy analyses with 720K genotypes. There has been concern, though mainly anecdotal, that APY would work less well for multi-breed populations or at least require larger core sizes than for single breeds. A simulation study by Vandenplas *et al.* (2018) demonstrated good performance of APY for crossbred data when the core, of size equal to the number of eigenvalues explaining 98 to 99% of variation in $\mathbf{G}$, included animals from all breed compositions. Dealing with a beef cattle population involving 41

**Table 2. Number of non-zero elements in $\mathbf{H}^{-1}$ (half-stored) for different APY schemes and number of iterates required to solve the corresponding mixed model equations**

| Select.[a] | Number of non-zero elements[b] | | | | | Number of PCG iterates | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2.5[c] | 5 | 10 | 15 | 20 | 2.5 | 5 | 10 | 15 | 20 |
| RND | 164 | 189 | 228 | 255 | 269 | 644 | 629 | 639 | 641 | 632 |
| PRG | 154 | 181 | 228 | 261 | 270 | 627 | 650 | 713 | 779 | 745 |

[a] Selection of core animals: RND random, PRG most progeny [b] In millions [c] Number of core animals; in 1000

**Table 3. Correlations between total predicted breeding values from single-step analyses using the 'full' inverse of the genomic relationship matrix and its TBLUP approximation**

| | Proportion of variation explained | | | | | |
|---|---|---|---|---|---|---|
| | 90% | 95% | 96% | 97% | 98% | 99% |
| No. of eigenvalues | 9,946 | 13,077 | 13,990 | 15,094 | 16,502 | 18,908 |
| No. of PCG iterates | 614 | 629 | 636 | 641 | 663 | 686 |
| Non-genotyped animals | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Genotyped animals | 0.9950 | 0.9982 | 0.9987 | 0.9991 | 0.9995 | 0.9998 |

breeds, Mäntysaari *et al.* (2017) recommended TBLUP as a well defined and automatic approach to approximate $\mathbf{G}^{-1}$ for any population structure. Our results suggest that approximation of $\mathbf{G}^{-1}$ using either APY or TBLUP can result in predicted breeding values which are virtually identical to those obtained inverting $\mathbf{G}$ directly, whilst offering the scope for reducing computational requirements. Details will depend on the implementation of ssGBLUP and have not been considered in this study; see Mäntysaari *et al.* (2017) for some discussion of respective strategies and timings. A suitable APY core size or number of PCs to be used for TBLUP was identified to be about 15K. This fell well within the range of corresponding values reported in the literature for single breed studies. However, current levels of genotyping for our data were relatively modest and, moreover, the distribution of genotypes over breeds was very uneven. It remains to be seen whether such levels of approximation will be representative as the number of genotypes increases, especially for the minor breed groups.

## CONCLUSIONS

Techniques available to approximate the inverse of the genomic relationship matrix in single step genomic evaluation can yield predicted breeding values for multi-breed sheep that are highly correlated with those obtained using a full inverse. Future work will need to re-evaluate suitable levels of approximation as numbers and breed diversity of genotyped animals increase.

## ACKNOWLEDGEMENTS

This work was supported by Meat and Livestock Australia grant L.GEN.1704.

## REFERENCES

Lourenco D.L., Tsurata S., Fragomeni B.O., Masuda Y., Aguilar I., Legarra A., Miller S., Moser D. and Misztal I. (2018) In *Proc. Eleventh World Congr. Genet. Appl. Livest. Prod.* Auckland, February 11–16. Paper No. 495.

Mäntysaari E.A., Evans R.D. and Strandén I. (2017) *J. Anim. Sci.* **95**:4728.

Masuda Y., Misztal I., Legarra A., Tsuruta S., Lourenco D.A.L., Fragomeni B.O. and Aguilar I. (2017) *J. Anim. Sci.* **95**:49.

Meyer K. (2007) *J. Zhejiang Univ. SCIENCE B* **8**:815.

Misztal I., Legarra A. and Aguilar I. (2014) *J. Dairy Sci.* **97**:3943.

Nilforooshan M.A. and Lee M. (2019) *J. Anim. Sci.* **97**:1090.

Pocrnic I., Lourenco D.A.L., Masuda Y., Legarra A. and Misztal I. (2016a) *Genetics* **203**.

Pocrnic I., Lourenco D.A.L., Masuda Y. and Misztal I. (2016b) *Genet. Sel. Evol.* **48**:82.

Strandén I., Matilainen K., Aamand G. and Mäntysaari E.A. (2017) *J. Anim. Breed. Genet.* **134**:264.

Van Raden P.M. (2008) *J. Dairy Sci.* **91**:4414.

Vandenplas J., Calus M.P.L. and ten Napel J. (2018) *J. Anim. Sci.* **96**:2060.

Vitezica Z.G., Aguilar I., Misztal I. and Legarra A. (2011) *Genet. Res.* **93**:357.