# COMPARING GENOMIC RELATIONSHIP MATRICES WITH RELATIONSHIP ESTIMATED FROM PEDIGREE DATA

**M.M. Farah[1], M.R.S. Fortes[2], L.R. Porto-Neto[3], C.T. Meira[1], M. Kelly[2], L.O.C. Duitama[1], A.V. Pires[4], R. Fonseca[1], S.S. Moore[2]**

[1]São Paulo State University, Jaboticabal, SP, Brazil. [2]The University of Queensland, Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, Brisbane, QLD 4062. [3]CSIRO Food Futures Flagship, and Division of Animal, Food and Health Science, Queensland Bioscience Precinct, Brisbane, QLD 4067. [4]Federal University of the Jequitinhonha and Mucuri Valleys, Diamantina, MG, Brazil

## SUMMARY

In this study, we tested 3 methods of building variations of genomic relationship matrix and compared these with the classic **A** matrix (pedigree based). Brahman bulls (n = 1,695) genotyped for or imputed to more than 700,000 single nucleotide polymorphisms were used. The allele frequencies used to obtain the 3 variations of **G** were: 0.5 for all SNPs (**G50**), the average minor allele frequency (**GMF**), and the observed allele frequency of each SNP (**GOF**). Our results indicate that, it is relevant to evaluate the allele frequency in the population and select the method of building matrices to increase the importance of rare alleles, which can help with estimating more precise relationships.

## INTRODUCTION

The use of genomic information has been growing in animal breeding programs. Several researchers are using this type of information to improve the accuracy of estimated breeding values (Hayes *et al.* 2010; Gianola *et al.* 2010; Erbe *et al.* 2012). Technology advancement and the possibility of genotyping many individuals made possible to use information on the alleles identical by state (IBS), not only identical by descendent (IBD) that can be shared through common ancestors. It is feasible to use a genomic relationship matrix (**G**) for estimating breeding values (Meuwissen and Goddard 1996).

Often it is not possible to obtain genomic information on the whole population and generate a relationship matrix based entirely on genomics due to the cost of genotyping and lack of samples to genotype. It is also difficult to estimate the allele frequencies of the base population. In the absence of this information, methods were developed to calculate the genomic relationship matrix using either an observed allele frequency, or a minor allele frequency or even a fixed value for allele frequency. These methods use observations from the genotyped population, which may be observed by actual genotyping or inferred with imputation from low density panels to high density panels. Forni *et al.* (2011) used genomic information from a population of pigs (1,919 females and 70 males) to test the impact on breeding values of using different approaches to build the **G** matrix and compared it to the **A** matrix (pedigree based). Forni *et al.* (2011) concluded that the breeding values estimated using the traditional **A** or an **H** matrix, that have both genomic and pedigree information, were similar. Their evidence suggested no real benefit from including genomic information in pig breeding programs. However, population structures in commercial pig lines are very different from the breed structure encountered in the beef cattle industry. Therefore, it is important to evaluate the contribution of genomic information to genetic evaluation processes in beef cattle. In this study, we test 3 methods of building genomic relationship matrices and compare these with the classic **A** matrix.

## MATERIALS AND METHODS

**Animals and genotypes.** Data from 1,695 Brahman bulls were used in the current study. These cattle represent a subset of the population bred by the Beef CRC that was described in detail previously (Burns *et al.* 2013; Corbet *et al.* 2013). This population has information on 729,068 single nucleotide polymorphisms (SNPs). These SNPs were genotyped (97 animals) or imputed (1,598 animals) from a lower density Illumina chip (BovineSNP50). Although only 97 animals within this study were genotyped on the high density marker panel, the full reference used to impute genotypes contained 917 animals from the Beef CRC population. Quality control criteria excluded SNP if minor allele frequency was lower than 0.05. Also, if pairwise correlations between SNP alleles was stronger than 0.95, only one SNP of the pair remained in the analysis. After quality control, 569,620 SNPs were used to estimate **G**, as follows:

$$G = \frac{(M-P)(M-P)'}{2\sum_{j=1}^{m} p_j(1-p_j)} \ ,$$

where **M** is an allele-sharing matrix with *m* columns (*m* = 569,620 SNPs) and *n* rows (*n* = 1,695 individuals), and **P** is a matrix containing the frequency of the second allele ($p_j$), expressed as $2p_j$. $\mathbf{M}_{ij}$ was 0 if the genotype of individual $i$ for SNP $j$ was homozygous for the first allele, was 1 if heterozygous, or 2 if the genotype was the other homozygous state.

The frequencies used to obtain 3 variations of **G** were similar to the methods described by Forni *et al.* (2011) where **P** matrix was obtained with: 1) the observed allele frequency of each SNP for the population (**GOF**), 2) average minor allele frequency (**GMF**), and 3) 0.5 for all SNPs (**G50**). For comparison between these variations of **G** matrices and the **A** matrix two methods were used: descriptive statistics and the correlation between the estimated kinship of individuals. For this population, 7 generations pedigree was used to obtain the relationship between the genotyped animals, underpinning the A matrix (total number of animals 3030). The comparison between **A** and **G** variations was made using only the relationship estimated between genotyped individuals.

## RESULTS AND DISCUSSION

Descriptive statistics for the **A** relationship matrix and the **G** relationship matrices, estimated for genotyped animals are provided in Table 1. The diagonal and off-diagonal elements were most similar for the matrices **A** and **GOF**, but the variances were very different. This lead to the differences between the matrices that can also be observed in Figure 1.

**Table 1. Statistics of relationship coefficients estimated using pedigree and genomic data***

| | Diagonal elements | | | | Off-diagonal elements | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Min. | Max. | Var. | Mean | Min. | Max. | Var. |
| **A** | 1.00 | 1.00 | 1.12 | $3.7\text{x}10^{-5}$ | 0.01 | 0.00 | 0.62 | $1.4\text{x}10^{-3}$ |
| **GOF** | 1.03 | 0.90 | 1.26 | $3.4\text{x}10^{-3}$ | 0.00 | -0.11 | 0.66 | $1.9\text{x}10^{-3}$ |
| **GMF** | 2.84 | 2.57 | 3.08 | $3.6\text{x}10^{-3}$ | 1.91 | 1.55 | 2.58 | $5.7\text{x}10^{-3}$ |
| **G50** | 1.36 | 1.20 | 1.52 | $1.4\text{x}10^{-3}$ | 0.68 | 0.45 | 1.16 | $2.6\text{x}10^{-3}$ |

***A** (relationship matrix pedigree-based); **GOF** (genomic relationship matrix with observed frequency); **GMF** (genomic relationship matrix with averaged minor allele frequency); **G50** (genomic relationship matrix with frequency 0.5 for all alleles).

Differences between the estimates for kinship based in either **A** or **G** calculations were observed (Figure 1). For some pairs of animals that **A** estimated as having no relationship (a value of zero), **G** matrices estimated values higher than zero suggesting that some of these animals share alleles that are IBS, but may not be IBD, or they may have a common ancestor that was missing from the pedigree records.

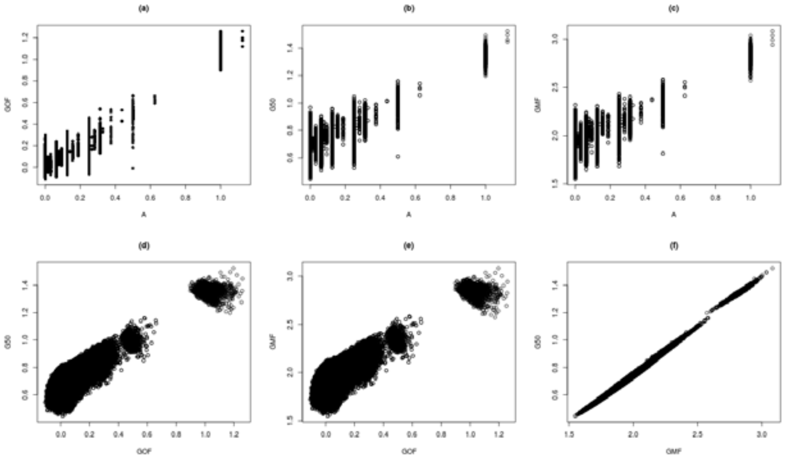

**Figure 1. Pairwise comparisons between kinship values estimated by matrices: a) A vs. GOF, b) A vs. GMF, c) A vs. G50, d) GOF vs. GMF, e) GOF vs. GMF, and f) GMF vs.G50.**

Lower variances for the matrix **A**, compared to **G**, can be explained because the method of calculating the relationship is by probability of two individuals sharing only alleles IBD. Higher variance among **G** elements, compared to **A**, can be expected because genomic relationships considered both alleles IBS and IBD, in agreement with Forni *et al.* (2011). Differences between matrix element variances are also reflected in the estimated correlations between **A** and **G** (Table 2). When the 3 variations of **G** were compared, we observed greater differences between **GOF** and the other 2 **G** matrices. The correlation between **G50** and **GMF** was high (Table 2), reflecting similar relationships estimated by these 2 methods (Figure 2).

**Table 2. Correlations between individual kinship estimates from each relationship matrix***

| | **GOF** | **GMF** | **G50** |
|---|---|---|---|
| **A** | 0.85 | 0.50 | 0.54 |
| **GOF** | | 0.58 | 0.63 |
| **GMF** | | | 0.99 |

***A** (pedigree based relationship matrix); **GOF** (genomic relationship matrix with observed allele frequencies); **GMF** (genomic relationship matrix with averaged minor allele frequencies); **G50** (genomic relationship matrix with a fixed allele frequency of 0.5 for all SNP).

The distribution of allele frequencies were shown in Figure 2. Extreme allele frequencies (higher than 0.8 and lower than 0.2) were observed in this population of Brahman bulls. This distribution of allele frequencies is in contrast with the distribution observed in pigs by Forni *et al.* (2011) that the distribution was much more uniform. The distribution of allele frequencies reflect the fact that the BovineSNP50 chip was developed primarily for *Bos taurus*. In *Bos taurus*, the allele frequencies are much more uniform in comparison to the distribution found in our Brahman population. The extreme frequencies in our allele distribution might help to explain the higher correlation between **GMF** and **G50**, and the differences between these 2 and **GOF**. Assumingly, if a fixed allele frequency such as 0.5 is used or minor allele frequencies are used instead of the observed frequencies, less importance is given to rare alleles and individual allele variation. As a result, **GMF** and **G50** might estimate higher values of kinship and some individuals might be perceived to be more related than suggested by **GOF** or **A** results. Differences in estimated kinship will influence estimated breeding values, having an impact on cattle selection programs.

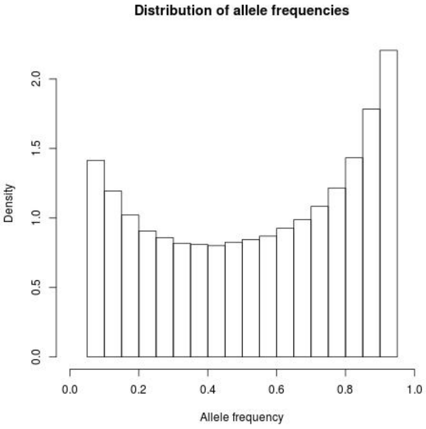

**Figure 2. Distribution of observed frequencies of the second allele.**

## CONCLUSION

In this study, relationships between individuals estimated using genomic data were correlated with estimates based on pedigree information. Since **G** matrices are correlated but not identical to the **A** matrix, genomic data can add information and contribute to accurate relationship estimations. Appropriate use of genomic information can be achieved with different methods of calculating a **G** matrix. Our results indicate that it is relevant to evaluate the allele frequency in the target population and select the methodology accordingly. Presence of extreme allele frequencies might indicate the importance of rare alleles and the use of **GOF**. Future work should test the influence of the different **G** matrices in the estimative of breeding values.

## ACKNOWLEDGEMENTS

The authors acknowledge that this research uses resources build by the Cooperative Research Centre for Beef Genetic Technologies (Beef CRC) and the financial support for genotyping Brahman animals was provided by Meat and Livestock Australia (project code B.NBP.0723). We thank the support of CAPES (Process: 13843/12-5). The Lab of scientific computation applied to animal science (LuCCA-Z) and QAAFI are acknowledge for providing the structure available.

## REFERENCES

Burns B.M, Corbet N.J., Cobert D.H., Crisp J.M., Venus B.K, Johnston D.J, Li Y, McGowan M.R and Holroyd R.G. (2013) *Anim Prod Sci* **53:**87.

Corbet N.J, Burns B. M, Johnston D. J, Wolcott M. L, Corbet D. H, Venus B. K, Li Y, McGowan M. R and Holroyd, R. G (2013) *Anim Prod Sci* **53**:101.

Erbe M., Hayes B. J, Matukumalli L. K, Goswami S, Bowman P. J, Reich C. M, Mason B. A and Goddard M. E (2012) *J Dairy Sci* **95**: 4114.

Forni S., Aguilar I., and Misztal I. (2011) *Genet Sel Evol* **43**:1.

Gianola D., Simianer H., and Qanbari S. (2010) *Genet Res (Camb)* **92**: 141.

Hayes B. J., Pryce J., Chamberlain A.J., Bowman P.J. and Goddard M. E. (2010) *PLoS Genet* **6**: e1001139.

Meuwissen T., and Goddard M. (1996) *Genet Sel Evol* **28**: 161.