Proc. Assoc. Advmt. Anim. Breed. Genet. 23:19-22

SIMPLE EXAMPLE TO DEMONSTRATE THE EFFECT OF ALLELE FREQUENCIES ON THE GENOMIC RELATIONSHIP MATRIX VALUES

M.H. Ferdosi, N.K. Connors, V. Boerner and D.J. Johnston

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

SUMMARY

Genomic evaluations using single-step genomic best linear unbiased prediction (ssGBLUP) combine the genomic relationship matrix (GRM) and numerator relationship matrix (NRM) together, to form the H matrix. The GRM values represent relationships between individuals and are dependent on allele frequencies. In this study, a simple example is used to demonstrate how the change in allele frequency can effect the values in the GRM, while also exploring the possible range of GRM values.

INTRODUCTION

In the pre-genomic era, pedigree was used to build the Numerator Relationship Matrix (NRM) that shows the relationship among individuals. The NRM is double the coancestry and can only show the relatedness between individuals, so the NRM values are always positive and range between 0 to 2. The NRM is a key component in Mixed Model Equations (MME) to calculate variance components and Estimated Breeding Values (EBVs). Genomics is used routinely in genetic evaluations nowadays, such as Australia's national beef recording and genetic evaluation system (BREEDPLAN), and with decreasing prices of genotyping, large numbers of individuals are genotyped. VanRaden (2008) showed that a Genomic Relationship Matrix (GRM) can replace the NRM in MME. The GRM is a variance and covariance matrix that can not only show relatedness among individuals but can also show the unrelatedness among individuals through negative values. The GRM values are dependent on allele frequencies and coding (Strandén and Christensen 2011; Tier *et al.* 2015). In the situation that both genotype (GRM) and pedigree (NRM) are available as current and historical information, a new method is required to make best use of both information sources appropriately. Single-Step genomic best linear unbiased prediction (ssGBLUP) was suggested by (Aguilar *et al.* 2010) to address this issue by building the new matrix H, combining both NRM and GRM information. Currently, ssGBLUP used in BREEDPLAN uses realised population allele frequencies to build the GRM. In this study, a simple example is used to demonstrate how the change in allele frequency can change the GRM values, whilst also exploring the possible range of GRM values. A better understanding of effects of allele frequency on GRM values will lead to a better understanding of the H matrix.

MATERIAL AND METHODS

Theory. This study considers a very simple situation where we have three animals, each with one marker (alleles AA, AB and BB). Summarising the GRM value (r) for one locus and two individuals using VanRaden first method (VanRaden 2008):

$$r = \frac{(b-2p+1)(c-2p+1)}{2p(1-p)} = \frac{bc+b+c+1}{2p(1-p)} - \frac{b+c+2-2p}{1-p} \quad (1)$$

where 'b' and 'c' were genotypes (only one marker) for two individuals and 'p' was the allele frequency. The 'b' and 'c' are coded -1, 0 and 1 for AA, AB and BB. The $(2p-1)$ that is subtracted from 'b' and 'c' is the mean genotype score. The $2p(1-p)$ is a scaling factor in order to make the GRM values comparable to NRM. For the case where 'b' and 'c' are opposing homozygotes i.e. b =

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

-1 and $c = 1$ then 'r' is

$$r = \frac{-1-1+1+1}{2p(1-p)} - \frac{-1+1+2-2p}{1-p} = 0 - \frac{2-2p}{1-p} = -2. \quad (2)$$

For the case where 'b' and 'c' are both heterozygote i.e. $b = 0$ and $c = 0$ then

$$r = \frac{1}{2p(1-p)} - \frac{2-2p}{1-p} = \frac{1}{2p(1-p)} - 2. \quad (3)$$

Table 1 - (A) shows the formulas for all genotype pairs and Table 1 - (B) shows similar formulas prior to dividing the GRM values by the scaling factor $2p(1-p)$. The determinant for both matrices were equal to 0, i.e. this matrix is singular and cannot be inverted as mentioned in Strandén and Christensen (2011). Table 2 shows the formula for which allele frequency can be calculated if wanting to obtain a specific relationship value. A relationship cannot be calculated for opposing homozygotes by using Table 1 - (A) when scaling factor is used, and as such there is no formula for this combination in Table 2. However, without the scaling factor (Table 1 - (B)) or changing the scaling factor the relationship can be calculated.

Table 1. Formula to calculate GRM value (r) for all possible genotype pairs - single marker only

Formula	(A) - with division by $2p(1-p)$			(B) - without division by $2p(1-p)$		
Allele	-1	0	1	-1	0	1
-1	$\frac{2p}{1-p}$	$\frac{2p-1}{1-p}$	-2	$4p^2$	$4p^2-2p$	$4p^2-4p$
0		$\frac{1}{2p(1-p)}-2$	$\frac{-2p+1}{p}$		$4p^2-4p+1$ or $(1-2p)^2$	$4p^2-6p+2$
1			$\frac{2}{p}-2$			$4p^2-8p+4$ or $(2-2p)^2$

p is the allele frequency

Table 2. Formula to calculate allele frequency (p) based on the specific relationship (r) in GRM - single marker only

Allele	-1	0	1
-1	$\frac{r}{r+2}$	$\frac{r+1}{r+2}$	?
0		$\frac{\sqrt{r^2+2r}+r+2}{2(r+2)}-2$	$\frac{1}{r+2}$
1			$\frac{2}{r+2}$

r is the relationship

RESULTS AND DISCUSSION

The formulas shown in Table 1 - (A) were used to calculate the GRM values that would be generated when p is 0.5 and 1. Table 3 - (A) shows the GRM values when p is 0.5, and Table 3 - (B) shows the GRM values when p is 1. Since the $2p(1-p)$ becomes 0 when the p value is 1, the limit was used when p approaches 1 (or 0 - Table 3 - (B)). Table 2 can be useful for simulation purposes. For example, Tables 4 (A) and (B) show the allele frequencies required to get a GRM values for important relationships of 0.5 (expected value for parent and offspring relationships or full-sib relationships) and 0.25 (expected value for half-sibs relationships) respectively. Figure 1 summarises the results shown in Tables 3 and 4.

Table 3. **Table shows GRM values when $p = 0.5$ (A) and when p approaches 1 (B) - by using the formula in Table 1 - A**

Formula	(A) - $p = 0.5$ and $2p(1-p) = 0.5$			(B) - $lim_{p->1+}$ and $2p(1-p) = 0$		
Allele	-1	0	1	-1	0	1
-1	1/0.5	0	−1/0.5	0	−1	−2
0		0	0		∞	∞
1			1/0.5			∞

Table 4. **For different relationships (r) using formula in Table 2 the p would be**

Formula	(A) - for 0.5 relationships			(B) - for 0.25 relationships		
Allele	-1	0	1	-1	0	1
-1	1/5	3/5	–	1/9	5/9	–
0		$-\frac{\sqrt{5}-5}{10}, \frac{\sqrt{5}+5}{10}$	2/5		1/3,2/3	4/9
1			4/5			8/9

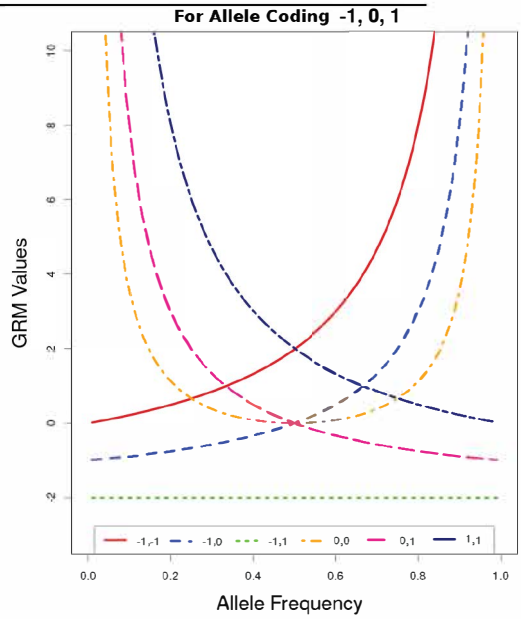

Figure 1. **Effect of different allele frequencies on the GRM values using three individuals and one locus. The legend shows the genotypes pairs.**

For a single marker only GRM, as discussed in this article, allele frequencies have significan effects on the GRM values. As shown in Figure 1, the more extreme the allele frequency (i.e. (or 1) the more extreme the GRM value. Table 3 - (B) shows that allele frequencies of 0 and 1 car result in infinite GRM values, demonstrated also in Figure 1. The lower limit of GRM for opposing homozygote is always -2, regardless of allele frequency. Figure 1 demonstrates how rare alleles and extreme allele frequencies can cause very large numbers in the GRM. This is amplified here due to only using a single marker. It should be noted that in practice, usually using thousands of markers

the effect of extreme allele frequencies will be minimized. This is dependent on SNP selection and whether the population is multi-breed for example. This simple example shows the importance of choosing the appropriate allele frequency (e.g. base population allele frequency – VanRaden (2008)) in order to reflect the true relationship among individuals in a GRM. Removing SNPs with very high or low allele frequencies or replacing their allele frequencies with pre-set allele frequencies may lead to more compatible values in GRM (in comparison to NRM), with no or negligible effect on estimated breeding values kings (Tier *et al.* 2015).

CONCLUSIONS

In this article a simplified version of the GRM was presented to demonstrate the effect of allele frequency on GRM values. In addition, simple formula were presented to calculate GRM values based on the specific allele frequency, or what allele frequency to use to obtain a specific GRM relationship value. These formulas can further be used for simulation purposes and development of methods to build the GRM efficiently.

ACKNOWLEDGMENTS

This research is supported by Meat and Livestock Australia (MLA) project L.GEN.0174. . The authors also wish to thank the reviewers for comments leading to the improvement of this article.

REFERENCES

Aguilar I., Misztal I., Johnson D., Legarra A., Tsuruta S. and Lawlor T. (2010) *J Dairy Sci* **93**:743.
Strandén I. and Christensen O.F. (2011) *Genet Sel Evol* **43**:25.
Tier B., Meyer K. and Ferdosi M. (2015) In *Proc. Assoc. Advmt. Anim. Breed. Genet*, 22. pp. 28–30
VanRaden P.M. (2008) *J Dairy Sci* **91**:4414.