# COMPARISONS OF IDENTICAL BY STATE AND IDENTICAL BY DESCENT RELATIONSHIP MATRICES DERIVED FROM SNP MARKERS IN GENOMIC EVALUATION

**S.A. Clark[1,2], B.P Kinghorn[1] and J.H.J. van der Werf[1,2]**

[1] University of New England, Armidale, NSW 2351
[2] CRC for Sheep Industry Innovation, University of New England, Armidale, NSW 2351

## SUMMARY

In animal populations, family members inherit alleles through common ancestors and these shared regions are referred to as identical by descent (IBD). Furthermore, animals may also share alleles due to a random association with no known common inheritance pattern. This identity by state (IBS) also covers known relationships, such that regions that are IBD are also IBS, however regions that are IBS may not always be IBD. In the genetic evaluation of livestock, IBD and IBS information can be used to build the genomic relationship matrix (GRM) and breeding values can be predicted using genomic best linear unbiased prediction (gBLUP).

This study compares a number of different methods to construct the GRM, using IBD and IBS information. Each method was evaluated using a reference dataset of 1781 Merino sheep and validated using 164 progeny tested sires that had accurate breeding values. Estimates of variance components were also compared. There was no significant difference between the accuracy achieved by the IBS and IBD methods. However the accuracy of the EBVs decreased as a greater restriction was applied to whether a region was IBD or not IBD. Furthermore, estimates of variance components were substantially different for IBD and IBS methods.

## INTRODUCTION

In animal populations there is often a high resemblance between the phenotypes of family members due to genes inherited from common ancestors (Fisher 1918). This theory has been widely discussed in the field of quantitative genetics and is currently used for the prediction of merit in livestock and detection of disease in humans (Henderson 1975; Donnelly 1983). In livestock genetic evaluation, best linear unbiased prediction (BLUP) (Henderson 1975) uses this concept to form the co-variances among the phenotypes of known relatives through the use of a numerator relationship matrix (NRM). Included in this matrix are coefficients of relationships which are the expected proportion of alleles that individuals share in common, identical by descent (IBD) based on pedigree information. Theories and methods using the same principles have also been described for the estimation of variance components.

Marker information has already been included in mixed model analyses (BLUP) using a relationship matrix derived from these markers, called the genomic relationship matrix (GRM) (Visscher *et al*. 2006; VanRaden 2008). This matrix can potentially describe the underlying covariance structure among individuals more fully than a matrix based on pedigree information alone, because the GRM uses estimates of realised relationships rather than expected relationships (Hayes *et al*. 2009). Popular methods for forming the genomic relationship matrix have been described by VanRaden (2008) and Yang *et al*. (2010). These methods use identical by state (IBS) information which is scaled by the allele frequencies to build the GRM, as shared rare alleles are more likely to be IBD than common alleles. However, these methods do not explicitly differentiate between IBD and IBS information. In contrast, there are very few methods that explicitly define IBD and often these methods only perform as well as IBS methods. However, many of these IBD methods have only been used in simulation and therefore constrained by the model used to simulate variation (Calus *et al*. 2008; Hickey *et al*. 2013).

In human quantitative genetics there has been a large focus placed on IBD information (Thompson 2008). Often, genotype probabilities that accommodate the probabilities of cross over events are used for determining IBD between individuals (Donnelly 1983) i.e. the more distant the relationship between two individuals the higher the probability that many crossovers have occurred. Many methods and programs have been described for the estimation of IBD and have been used for the detection of regions of the genome that are IBD, e.g. PLINK (Purcell *et al*. 2007) and fastIBD (Browning and Browning 2011).

The aim of this study was to compare the use of IBD and IBS genomic relationship information to predict genomic breeding values using real data. The differences between each GRM were investigated, together with their effects on the estimation of breeding values and variance components, and the accuracies of resulting estimates of breeding value (EBVs).

**METHODS**

The data used in this study consisted of phenotypic and genotypic records from the Australian Sheep Cooperative Research Centre (CRC) information nucleus flock (INF). This dataset consisted of a reference dataset consisted of phenotypic and genotypic records from 1781 merino animals and a validation dataset of 164 merino sires with accurate Australian Sheep Breeding Values (ASBV's). Definitions of ASBV's can be obtained from Sheep Genetics Australia. Phenotypic information on the trait scanned eye muscle depth (SEMD) was analysed. To observe the effect of relatedness, the validation population was split into three groups based on their pedigree relationship to the animals in the reference dataset (Clark *et al*. 2012). The three groups consisted of; 50 closely related animals (Close), with a maximum relationship of greater than 0.25; the 54 distantly related animals (0.01-0.249) (Dist); and 60 unrelated animals that shared zero pedigree relationship.

All animals in each dataset were genotyped using the Illumina 50K ovine SNP chip. All SNP in this dataset underwent a number of genotyping quality control measures (see Daetwyler *et al*. (2010)). The following fixed effects were fitted in the analysis of SEMD: Sex, birth type, rearing type, age of dam, contemporary group (birth year • birth month) (site • management group), age-at-trait recording and live weight at scanning.

As in Daetwyler *et al*. (2010) we assumed the gBLUP model;

$$y = Xb + Zg + e$$

where y is a vector of phenotypes, X is a design matrix relating the fixed effects (as described above) to each animal, b is a vector of fixed effects, Z is a design matrix allocating records to breeding values, g is a vector additive genetic effects for animals in the reference dataset and the validation dataset and e is a vector of random normal deviates $\sigma_e^2$. Furthermore V(g) = G$\sigma_g^2$ where G is the genomic relationship matrix, and $\sigma_g^2$ is the genetic variance for this model. The GRM (G) was formed using two IBS methods defined by ($G_V$) VanRaden (2008) and ($G_Y$) Yang *et al*. (2010) and five IBD methods were also evaluated. Two fastIBD matrices (Browning and Browning 2011) were formed. $G_{Fast(h)}$ was based on the stringent threshold for IBD used in human genetics and $G_{Fast(R)}$ used a relaxed threshold on whether a region was IBD or not. Three probability of IBD methods (Kinghorn 2012) were also used: $G_{Prob}$ used an IBD probability estimate for each individual loci that was based on IBD information from adjacent marker information. This method was extended such that regions were identified as IBD if animals shared haplotypes of 10 ($G_{Prob10}$) and 50 ($G_{Prob50}$) markers with an IBD probability of greater than 0.98 and if regions were shorter than the given length they were assumed to be IBS and did not contribute to the estimate of relationship.

## RESULTS

The IBS and unrestricted IBD methods ($G_{Prob}$ and $G_{Fast(R)}$) were the most accurate methods to predict breeding value (Table 1). These results are similar to simulation studies by Hickey *et al.* (2013) and Calus *et al.* (2008) were there was little difference between the IBS and best IBD methods. However, accuracy was reduced when a restriction was placed on whether a region was IBD or not, by either increasing the length of the IBD segment as in $G_{Prob10}$ and $G_{Prob50}$ or by increasing the significance threshold as in $G_{Fast(h)}$. The highly restricted fastIBD ($G_{Fast(h)}$) method was the least accurate method (Table 1).

**Table 1 The EBV accuracy (correlation between gEBV and ASBV) and regression of gEBV on ASBV, estimated using the alternative ways to define the genomic relationship matrix**

| | **IBS**[*] | | **IBD**[i] | | | | |
|---|---|---|---|---|---|---|---|
| | $\mathbf{G_V}$ | $\mathbf{G_Y}$ | $\mathbf{G_{Prob}}$ | $\mathbf{G_{Prob10}}$ | $\mathbf{G_{Prob50}}$ | $\mathbf{G_{Fast(R)}}$ | $\mathbf{G_{Fast(h)}}$ |
| **Accuracy** | | | | | | | |
| All Animals | 0.456 | 0.451 | 0.453 | 0.413 | 0.340 | 0.465 | 0.239 |
| Unrelated | 0.224 | 0.206 | 0.226 | 0.226 | 0.172 | 0.281 | 0.137 |
| Distantly related | 0.450 | 0.499 | 0.478 | 0.394 | 0.334 | 0.434 | 0.216 |
| Closely related | 0.640 | 0.643 | 0.650 | 0.622 | 0.555 | 0.668 | 0.413 |
| **Regression** | | | | | | | |
| All Animals | 0.882 | 0.873 | 0.914 | 1.033 | 1.249 | 1.011 | 0.834 |

[*] IBS methods were constructed using methods by VanRaden ($\mathbf{G_V}$) and Yang *et. al.* ($\mathbf{G_Y}$)

[i] IBD methods were constructed using: 1) IBD probabilities ($\mathbf{G_{Prob}}$) with different haplotype lengths ($\mathbf{G_{Prob10}}$ and $\mathbf{G_{Prob50}}$) and 2) the FastIBD module of the Beagle software ($\mathbf{G_{Fast}}$) with either a relaxed ($\mathbf{G_{Fast(R)}}$) or strict ($\mathbf{G_{Fast(h)}}$) constraint on whether a region was IBD or not.

When animals were unrelated or distantly related to the reference population, accuracy was reduced for both IBD and IBS methods. Accuracy decreased in all cases when the IBD segment length increased. Furthermore, when fast IBD was highly restricted ($G_{Fast(h)}$), its ability to predict breeding value in unrelated animals was also reduced. A reduction in accuracy was observed because, as the restriction on whether a region was IBD or not increased, some useful information about rare, short haplotypes was lost. Interestingly, in unrelated animals, the $G_{Fast(R)}$ tended to be the most accurate method (although not significantly better than $G_V$, $G_Y$ or $G_{Prob}$).

Table 1 also shows the regression of GEBV on ASBV for each of the different GRMs. It shows that the IBS and $G_{Prob}$ methods had a regression coefficient less than one, showing the GEBVs are over dispersed. In contrast, the $G_{Fast(R)}$ and $G_{Prob(10)}$ methods had a regression coefficient close to 1 showing that both sets of EBV's are on a similar scale to the progeny tested ASBV's. The IBS methods: $G_V$ and $G_Y$ are very similar and resulted in a 0.999 correlation between the breeding values estimated using these methods. The $G_{Fast(R)}$ and $G_{Prob}$ methods were only slightly different with a correlation between breeding values of 0.96 and 0.94 respectively with the IBS methods. Finally, $G_{Prob(10)}$ used partially different information as the breeding values estimated from this method were only 0.88 correlated with $G_V$. Although the methods appear to be very similar, given the high correlation between breeding values, the variance components (Table 2) estimated from each method were different.

The $G_V$ and $G_Y$ methods by VanRaden (2008) and Yang *et al.* (2010) resulted in similar variance component estimates. The ProbIBD methods ($G_{Prob}$ and $G_{Prob10}$) also resulted in higher estimates of genetic variance. In contrast, the fastIBD method ($G_{Fast(R)}$) resulted in a substantially lower estimate of genetic variance and therefore heritability. This implicitly shows that the scale of the various GRM's (which relates to the methods used to construct each GRM) can have a large impact on variance component estimation.

**Table 2 Variance components estimated using various methods to define the genomic relationship matrix**

| | Pedigree | IBS | | IBD | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $G_V$ | $G_Y$ | $G_{Prob}$ | $G_{Prob10}$ | $G_{Prob50}$ | $G_{Fast(R)}$ | $G_{Fast(H)}$ |
| Va | 1.12 | 1.288 | 1.305 | 1.883 | 1.545 | 1.636 | 0.715 | 0.904 |
| Ve | 3.03 | 3.021 | 3.015 | 2.935 | 2.778 | 2.646 | 3.635 | 3.387 |
| Vtotal | 4.15 | 4.309 | 4.320 | 4.818 | 4.323 | 4.282 | 4.350 | 4.291 |
| $h^2$ | 0.269 | 0.299 | 0.302 | 0.391 | 0.357 | 0.382 | 0.16 | 0.211 |

[*] IBS methods were constructed using methods by VanRaden (**$G_V$**) and Yang *et. al.* (**$G_Y$**)

[i] IBD methods were constructed using: 1) IBD probabilities (**$G_{Prob}$**) with different haplotype lengths (**$G_{Prob10}$** and **$G_{Prob50}$**) and 2) the FastIBD module of the Beagle software (**$G_{Fast}$**) with either a relaxed (**$G_{Fast(R)}$**) or strict (**$G_{Fast(h)}$**) constraint on whether a region was IBD or not.

## CONCLUSION

This study shows that IBD probabilities and information from the fastIBD module of Beagle can be used to predict breeding value in real data. Furthermore, this study has shown that some IBD relationship matrices will perform as well as IBS based methods for genomic evaluation, even in unrelated animals. However, accuracy will reduce when breeding values are estimated using IBD methods that place a large restriction on whether a region is IBD or not. The variance components estimated from each GRM is impacted by the scale of the relationship matrix. The scale is impacted by the definition of the relationship information (IBD or IBS) and the allele frequencies that are used to scale the GRM.

## REFERENCES

Browning B.L. and Browning S.R. (2011). *Am. J. Hum. Genet.* **88**: 173.

Calus M.P.L., Meuwissen T.H.E., de Roos A.P.W. and Veerkamp R.F. (2008) *Genetics*. **178**:553.

Clark S.A., Hickey J.M., Daetwyler H.D. and van der Werf J.H.J. (2012). *Genet. Sel. Evol.* **44**: 4.

Daetwyler H.D., Hickey J.M., Henshall J.M. and Dominik S. (2010). *Anim. Prod. Sci.* **50**: 1004.

Donnelly K.P. (1983). *Theor. Pop. Biol.* **23**: 34.

Fisher R.A. (1918). *Trans. R. Soc. Edin.* **52**: 399.

Hayes B.J., Visscher P.M. and Goddard M.E. (2009). *Genet. Res.* **91**: 47.

Henderson C.R. (1975). *Biometrics* **31**: 423.

Hickey J.M., Kinghorn B.P., Teir B. Clark S.A. and van der Werf J.H.J (2013) *J Anim Breed Genet.* **130**: 259.

Kinghorn B.P. (2012). *Proc. Inter. Conf. Quant. Genet.* **32**:80.

Purcell S., Neale B., Todd-Brown K., Thomas L. and Ferreira M.A. (2007). *Am. J. Hum. Genet.* **81**:559.

Thompson E.A. (2008). *Theor. Pop. Biol*. **73**: 369.

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

Visscher P.M., Medland S.E., Ferreira M.A., Morley K.I. and Zhu, G. (2006). *PLoS Genet*. **2**:41.

Yang J., Benyamin B., McEvoy B.P., Gordon S.D., and Henders A.K. (2010). *Nat.Genet.* **42**:565.

**Table 2 Variance components estimated using various methods to define the genomic relationship matrix**

| | Pedigree | IBS | | IBD | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $G_V$ | $G_Y$ | $G_{Prob}$ | $G_{Prob10}$ | $G_{Prob50}$ | $G_{Fast(R)}$ | $G_{Fast(H)}$ |
| Va | 1.12 | 1.288 | 1.305 | 1.883 | 1.545 | 1.636 | 0.715 | 0.904 |
| Ve | 3.03 | 3.021 | 3.015 | 2.935 | 2.778 | 2.646 | 3.635 | 3.387 |
| Vtotal | 4.15 | 4.309 | 4.320 | 4.818 | 4.323 | 4.282 | 4.350 | 4.291 |
| $h^2$ | 0.269 | 0.299 | 0.302 | 0.391 | 0.357 | 0.382 | 0.16 | 0.211 |

[*] IBS methods were constructed using methods by VanRaden ($\mathbf{G_V}$) and Yang *et. al.* ($\mathbf{G_Y}$)

[i] IBD methods were constructed using: 1) IBD probabilities ($\mathbf{G_{Prob}}$) with different haplotype lengths ($\mathbf{G_{Prob10}}$ and $\mathbf{G_{Prob50}}$) and 2) the FastIBD module of the Beagle software ($\mathbf{G_{Fast}}$) with either a relaxed ($\mathbf{G_{Fast(R)}}$) or strict ($\mathbf{G_{Fast(h)}}$) constraint on whether a region was IBD or not.

## CONCLUSION

This study shows that IBD probabilities and information from the fastIBD module of Beagle can be used to predict breeding value in real data. Furthermore, this study has shown that some IBD relationship matrices will perform as well as IBS based methods for genomic evaluation, even in unrelated animals. However, accuracy will reduce when breeding values are estimated using IBD methods that place a large restriction on whether a region is IBD or not. The variance components estimated from each GRM is impacted by the scale of the relationship matrix. The scale is impacted by the definition of the relationship information (IBD or IBS) and the allele frequencies that are used to scale the GRM.

## REFERENCES

Browning B.L. and Browning S.R. (2011). *Am. J. Hum. Genet.* **88**: 173.

Calus M.P.L., Meuwissen T.H.E., de Roos A.P.W. and Veerkamp R.F. (2008) *Genetics*. **178**:553.

Clark S.A., Hickey J.M., Daetwyler H.D. and van der Werf J.H.J. (2012). *Genet. Sel. Evol.* **44**: 4.

Daetwyler H.D., Hickey J.M., Henshall J.M. and Dominik S. (2010). *Anim. Prod. Sci.* **50**: 1004.

Donnelly K.P. (1983). *Theor. Pop. Biol.* **23**: 34.

Fisher R.A. (1918). *Trans. R. Soc. Edin.* **52**: 399.

Hayes B.J., Visscher P.M. and Goddard M.E. (2009). *Genet. Res.* **91**: 47.

Henderson C.R. (1975). *Biometrics* **31**: 423.

Hickey J.M., Kinghorn B.P., Teir B. Clark S.A. and van der Werf J.H.J (2013) *J Anim Breed Genet.* **130**: 259.

Kinghorn B.P. (2012). *Proc. Inter. Conf. Quant. Genet.* **32**:80.

Purcell S., Neale B., Todd-Brown K., Thomas L. and Ferreira M.A. (2007). *Am. J. Hum. Genet.* **81**:559.

Thompson E.A. (2008). *Theor. Pop. Biol.* **73**: 369.

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

Visscher P.M., Medland S.E., Ferreira M.A., Morley K.I. and Zhu, G. (2006). *PLoS Genet*. **2**:41.

Yang J., Benyamin B., McEvoy B.P., Gordon S.D., and Henders A.K. (2010). *Nat.Genet.* **42**:565.