Proc. Assoc. Advmt. Breed. Genet. 21: 461-464

WHICH GENOMIC RELATIONSHIP MATRIX?

B. Tier[1], K. Meyer[1] and M. H. Ferdosi[1,2]

[1]Animal Genetics and Breeding Unit*, University of New England, Armidale NSW 2351.
[2]School of Environmental and Rural Science, University of New England, Armidale NSW 2351.

SUMMARY

Genomic information can accurately specify relationships among animals, including between those without known common ancestors. Genetic variances estimated with genomic data relate to unknown, more distant, founder populations than those defined by the pedigree. Starting from different sets of assumptions, the properties of some alternative genomic relationship matrices (**G**) are explored. Although the assumptions and matrices differ, the resulting sets of estimated breeding values predict the differences between animals identically, despite obtaining different estimates of the additive genetic variance – showing that there are many ways of building **G** that provide identical results. For some methods integer and logic, rather than floating point, operations will expedite building **G** many-fold.

INTRODUCTION

Genomic data can provide more accurate information about relationships among animals. When only pedigree information is available, progeny are assumed to receive a random half of each parents' genes and full-sibs are expected to share half their genes. With genomic data we can tell which half of each parents' genes an animal receives and precisely the proportion of genes shared by full-sibs. Generally, genomic information provides more detailed information about relationships including that between individuals that share no known common ancestors.

When a population is genotyped a genomic relationship matrix (**G**) takes the place of the numerator relationship matrix (**A**) in routine genetic analyses. However, unlike **A**, **G** must be built explicitly which can be a time consuming process particularly when the number of loci and/or genotyped animals is large. When $\mathbf{G}^{-1}$ is needed, **G** must also be inverted directly as it is dense and unlike **A**, **G** has no simple inverse. This operation is generally more computationally expensive than building **G** whereas $\mathbf{A}^{-1}$ can be constructed rapidly, directly from the pedigree.

Recently Forni *et al.* (2012) examined the effect of using different assumptions to build **G** but obtained the same results for some methods. This paper illustrates how using different assumptions when building **G**, can result in different **G** matrices and even estimated genetic variances, yet provide the same estimated breeding values (EBVs). It also shows how different assumptions can significantly expedite the process of building **G**.

THEORY

Estimates of relationships among individuals are essential for genetic evaluation. Traditionally **A** fulfilled that purpose. When combined with the genetic variance (σ_u^2), variance of the breeding values (**u**) was defined to be $\text{Var}(\mathbf{u})=\mathbf{A}\sigma_u^2$. **A** is based on the idea of identity by descent (IBD) and is built by tracing the flow of genes down the pedigree. Elements of **A** are twice the coancestry coefficient which are probabilities that limit the range of elements in **A** to [0,2]. Founders, the remotest set of ancestors with unknown pedigree, are assumed to be a random sample from a very large population in Hardy-Weinberg equilibrium. The partition of **A** relating to the founders is an identity matrix, which implies that the genome of each founder consists of two subsets. The first subset contains loci that are all homozygous and common to all founders and thus generate no

* AGBU is a joint unit of NSW Primary Industry and the University of New England

phenotypic variance. The other subset contains all loci that generate phenotypic variation. They are unique to each founder as off-diagonal elements of zero imply that there is no covariation with any other founder. This suggests that there were an infinite number of alleles at every locus in the base population.

Genomic data, in the form of single nucleotide polymorphisms (SNP), can be used to build **G** (Van Raden 2008) for individuals with genotypes. Using markers involves the strong assumptions relating to identity by state (IBS), where markers are deemed to be in linkage disequilibrium with genes affecting phenotypes, and that such genes behave similarly across the whole population, especially for relationships beyond the pedigree. When all individuals in the population have genotypes then **G** can be used in place of **A** so that the assumption about the variance of the breeding values becomes $\text{Var}(\mathbf{u})=\mathbf{G}\sigma_u^2$. A variety of different methods are available for building **G** and some of them are equivalent to including the SNP directly as individual effects (**g**) in the model (Stranden and Garrick, 2009) in place of the breeding values, so that $\mathbf{u}=\mathbf{Zg}$ and $\text{Var}(\mathbf{g})=\mathbf{I}\sigma_g^2$, where σ_g^2 is the variance due to the SNPs. The equivalence between these methods indicates a degree of ambiguity and loosely implies that the effects of the SNPs, or the quantitative trait loci in linkage disequilibrium with them are estimable. Some methods for building **G** result in elements that have no probabilistic interpretation (e.g. elements less than zero).

Genomic data. SNPs are the genotypes used in this paper, with each individual-locus represented by a number 0, 1 or 2, being the number of one of the alleles available at the locus. There are a animals with h haplotypes (h=2a) and m loci. The genotypes are represented by **Z**, an a x m matrix and haplotypes by **X** an h x m matrix**.** Haplotypes for each locus are formed independently of other loci. The matrix $\mathbf{K}=\mathbf{I}\otimes[1\ 1]$, where $\otimes$ is the Kronecker product, converts **X** to **Z** as **Z**=**KX**. The matrix **P** is conformable to **Z** and contains the allele frequencies (p) for each locus in its columns. In addition let **J** denote a matrix with all elements equal to 1. Dimensions of **J** are as implied in the equation where it is used. Where necessary we specify the row (i) and column (j) dimensions as subscripts ($\mathbf{J}_{ij}$).

G matrices. Three alternative methods for building **G** are considered. The first of these is Van Raden's (2008) first method, viz. $\mathbf{G}=\mathbf{MM}'/d$**,** where **M**=**Z**-2**P**, and $d=2\sum p(1-p)$. By subtracting 2**P** from **Z** genotypes are centred so that columns of **M** sum to zero. The denominator is designed to scale the matrix **G** to be similar to the scale of **A**. This formulation of **G** generates some irregular elements that cannot be interpreted as co-ancestry. These include negative elements, parent-offspring elements less than 0.5 and diagonals less than 1. Potentially, elements can be greater than 2 (between pairs of individuals sharing a very large number of low frequency alleles).

The second method is similar to the first with genotypes centred around zero: $\mathbf{F}=(\mathbf{Z}-\mathbf{J})(\mathbf{Z}-\mathbf{J})'/c$. The denominator, c, can be the same as d, or alternatively with all allele frequencies set to 0.5, c=m/2. **F** can also contain unusual elements, with the diagonal elements being a function of the proportion of the animals' loci that are homozygous. Elements of **F** are readily computed by counting the numbers of identical and of opposing homozygotes between each pair of animals. This allows the use of integer and logical operations that are much faster than floating point operations required to compute $(\mathbf{Z}-2\mathbf{P})(\mathbf{Z}-2\mathbf{P})'$.

The third method is based on building a gametic relationship matrix (**H**). Nominally, a gametic relationship matrix ($\mathbf{\Gamma}_i$) is built for each locus by counting 1 if the alleles are the same and 0 if they differ. Subsequently the complete gametic relationship matrix (**Γ**) is calculated by summing all the loci matrices and dividing by m. This is converted to the animal relationship as $\mathbf{H}=\mathbf{K\Gamma K}'/2$**.** In practice, **H** is built as $\mathbf{H}=\mathbf{K}[\mathbf{XX}'+(\mathbf{X}-\mathbf{J})(\mathbf{X}-\mathbf{J})']\mathbf{K}'/2m$. The method for building **H** ensures that it has no elements less than 0 nor greater than 2 and no diagonal elements less than 1.

Similarity. Expansion of the terms in the matrices illustrates the differences between them.

1. Considering **M** as **Z**-**J**-**D**, where **D**=2**P**-**J** the numerator of **G** (=**MM**'/d) gives

$\mathbf{MM}' = (\mathbf{Z}-\mathbf{J}-\mathbf{D})(\mathbf{Z}-\mathbf{J}-\mathbf{D})' = (\mathbf{ZZ}'-\mathbf{ZJ}'-\mathbf{ZD}'-\mathbf{JZ}'+\mathbf{JJ}'+\mathbf{JD}'-\mathbf{DZ}'+\mathbf{DJ}'+\mathbf{DD}')$

By setting $\mathbf{E} = -\mathbf{ZD}'+\mathbf{JD}'-\mathbf{DZ}'+\mathbf{DJ}'+\mathbf{DD}'$ and noting that with $\mathbf{J}_{am}\mathbf{J}_{am}'=m\mathbf{J}_{aa}$,

$\mathbf{G} = (\mathbf{ZZ}'+m\mathbf{J}-\mathbf{ZJ}'-\mathbf{JZ}'+\mathbf{E})/d$

2. $\mathbf{F} = (\mathbf{Z}-\mathbf{J})(\mathbf{Z}-\mathbf{J})'/c = (\mathbf{ZZ}'+m\mathbf{J}-\mathbf{ZJ}'-\mathbf{JZ}')/c.$
3. $\mathbf{H} = (\mathbf{K}[\mathbf{XX}'+(\mathbf{X}-\mathbf{J})(\mathbf{X}-\mathbf{J})']\mathbf{K}/2)/m$
   $= (\mathbf{KXX}'\mathbf{K}'+\mathbf{KJJ}'\mathbf{K}'/2-\mathbf{KXJ}'\mathbf{K}'/2-\mathbf{KJX}'\mathbf{K}'/2)/m$, and since $\mathbf{Z} = \mathbf{KX}$ and $\mathbf{KJ}_{hm} = 2\mathbf{J}_{am}$,
   $\mathbf{H} = (\mathbf{ZZ}'+2m\mathbf{J}-\mathbf{ZJ}'-\mathbf{JZ}')/m.$

These results clearly show how **G**, **F** and **H** differ and that since, $\mathbf{G} = (\mathbf{F}c+\mathbf{E})/d$ and $\mathbf{F} = m(\mathbf{H}-\mathbf{J})/c$, how one can be determined from another. When c=d, $\mathbf{G} = \mathbf{F}+\mathbf{E}/d$.

MATERIALS AND METHODS

A small population made up of four sires mated to the same five dams each producing one offspring was generated. Each individual had two haplotypes of 99 SNPs, a breeding value and phenotype for a trait with a heritability of 0.55. These were analysed with the model $\mathbf{y}=\mu+\mathbf{Z}_l\mathbf{u}+\mathbf{e}$, where the data are a function of the mean (μ), the breeding values ($\mathbf{u}$) and a residual ($\mathbf{e}$), and $\mathbf{Z}_l$ is an incidence matrix assigning observations to breeding values. $Var(\mathbf{u})=\mathbf{W}\sigma_u^2$, where **W** is a relationship matrix and $Var(\mathbf{e})=\mathbf{I}\sigma_e^2$. Genetic parameters for this population were estimated using five different matrices **W.** The first used **G** with a small amount (0.01**I**) added to make it invertible (positive definite), the second and third used **F** with denominators of d and m/2 respectively, the fourth used **H** and the last used **F**+10**J**. These data were analysed with WOMBAT (Meyer, 2007) to estimate variance components and breeding values.

RESULTS AND DISCUSSION

G matrices. The construction of the various matrices shows clearly how they relate to each other. The difference between **G** and **F**(c=m/2) arises from the different allele frequencies. **F** and **H** differ in their denominators and there is an additional term (m**J**) included in **H** that is not in **F**.

Evaluations. The results in Table 1 show that, regardless of which **W** matrix is used, the estimated breeding values (EBVs) are the same. The correlations between EBVs from different analyses are 1, or close to 1, as are the regressions of 1 on those obtained when **W**=**G**. Differences in estimated means are unimportant as EBVs are relative measures of genetic merit. Slight differences occur when **G** is used, compared to the other methods as its diagonal was augmented and some difference in the mean may be due to **E**. The addition of 10**J** to **F** has no effect, indicating that adding any multiple of **J** (results not shown) to these matrices have no effect. These results show the practice of augmenting the diagonal of **G** should be superseded by adding k**J**, where k is small, to ensure **G** is invertible. The likelihoods and residual variances are also the same for all models. Similar genetic variances were estimated when **G** or **F** was used. While the addition of a multiple of **J** to **F** matrices has no effect, it suggests a higher degree of relationship in that population than **F** alone. Using **H** obtained a considerably higher additive genetic variance

Table 1: **Results from evaluation of simulated data using different relationship matrices**

Relationship Matrix	Log-Likelihood	σ_e^2	σ_u^2	μ	Regression of EBVs on EBVs(**G**)		Correlation EBVs with EBV(**G**)
					Intercept	Slope	
G	-76.55	51.22	31.74	0.000	-	-	-
F(c=d)	-76.55	51.27	31.71	-0.027	0.027	0.999	1.0
F(c=m/2)	-76.55	51.27	33.83	-0.027	0.027	0.999	1.0
H	-76.55	51.27	67.73	-0.027	0.027	0.999	1.0
F+10**J**	-76.55	51.25	31.74	-0.027	0.027	0.997	1.0

than other matrices. This might suggest that **H** uses a more ancient set of founders than assumed when **G** or **F** is used. However, since $Var(\mathbf{u})=\mathbf{W}\sigma_u^2$, and if it is only their denominators that differ ($\mathbf{W}_1=w\mathbf{W}_2$), the estimated additive genetic variance must vary in a complementary manner ($\sigma_{u1}^2=\sigma_{u2}^2/w$). This is so for **F**(c=d) and **H** where the ratio of the additive genetic variances is d/m and similarly for **F**(c=d) and **F**(c=m/2) where this ratio is 2c/m.

Although the various genomic relationship matrices were different, their inverses, also necessarily different, provide the same results which may seem surprising given the different assumptions. Despite this, the same results indicate that the inverses are simple functions of each other showing that the genomic data are being used in exactly the same way.

The equivalence between these methods, based on relationship matrices, can be illustrated by considering modelling the genotypes directly. With this model the addition of a constant to the SNP genotypes for each locus has no effect on anything but the overall mean. The additive breeding values (**u**=**Zg**) would be the same as if nothing had been added. This is akin to centering alleles around different values and adding terms like **E** and k**J** to any **W**.

These results show that different approaches to using genomic data may not ensure real differences and may explain why some methods used by Forni *et al.* (2011) have identical results. These results also show that the apparent problems relating to strange elements (negative off-diagonals, and diagonals less than 1) in **G** are nothing to fear, they are simply on a different scale to the other **W**s. Starting with the idea of SNP similarity provides **H** which, by construction, can have a similar probabilistic interpretation to **A**. However, **H** provides a much greater genetic variance than the other methods, but this can be modified by factoring it by c/m.

As genomic data provide relationships among individuals that are not IBD, it is clear that the unknown founder population implied when genomic data are used must be different to the known founder population derived from pedigrees. These results show that the estimated additive genetic variance is sensitive to assumptions about allele frequencies which determine the denominator and, indirectly, the unknown founder population. Paradoxically, the EBVs estimated from each of these evaluations are insensitive to the different estimates of additive genetic variance when combined with the appropriate **W**. Conversely, incorrect EBVs could result from combining a relationship matrix **W** with an inappropriate additive genetic variance.

Building the numerators of **F** and **H** are based on **Z** and **X**. These matrices are integers and provide the opportunity to use integer rather than floating point operations. Furthermore, as the non-zero elements of **Z**-**J** are only 1, and -1 the process of building **F** can be done with logic operators which is magnitudes faster than the floating point operations used to build **G.**

CONCLUSION

Many ways of using genomic data to determine relationships among individuals in a population, while appearing to be different, are similar. Although they may be based on different assumptions, and can provide different estimates of the additive genetic variance, they provide the same measures of genetic merit of the population. The estimate of the additive genetic variance is sensitive to the estimate of allele frequencies. **F** should be used in place of **G,** as it is much quicker to build and provides an equivalent model and it does not require augmenting the diagonal to make it invertible.

REFERENCES

Forni, S., Aguilar, I., and Misztal I. (2011) *Genet. Sel. Evol*. **43**:1.
Meyer, K. (2007) *J. Zhejiang Univ. Sci. B* **8**:815.
Stranden I., Garrick D. (2009) *J. Dairy Sci*. **92**:2971.
Van Raden P.M. (2008) *J. Dairy Sci*. **91**:4414.