COMPARISON OF MEASURES OF RELATEDNESS USING PEDIGREE OR GENOMIC DATA IN A MULTI-BREED SHEEP POPULATION

B. Auvray and K.G. Dodds

AgResearch Ltd, Invermay Agricultural Centre, Mosgiel 9053, New Zealand

SUMMARY

Numerator relationship matrices (NRM) between individuals based on SNP genotypes, estimated using a method proposed by VanRaden (2008), and combined with modifications which rescale the NRM and which account for population substructure were compared with **A**, the NRM derived from pedigree. Getting matrices closely resembling the **A** matrix may be desirable because, in a crossbred or multi-breed context, the elements of **A** (particularly off-diagonal elements between breeds) are closer to the true average identity by descent between individuals. On a crossbred sheep data set of 7,855 individuals genotyped for 47,084 SNP, the NRM where population stratification was not accounted for performed poorly (overall mean absolute difference (MAD) from pedigree relatedness = 2.8%, MAD in Texel, the most differentiated breed in the dataset in term of allele frequencies, = 24.6%) while NRM corrected for population structure performed better (overall MAD = 1.2%, MAD in Texel = 2.9%). The impact of the rescaling was marginal, as it only reduced the overall and per breed MAD from pedigree by up to 0.1%.

INTRODUCTION

Many methods to estimate numerator relationship matrices (NRM) from SNP to use e.g. in genome enabled prediction (GEP, Meuwissen *et al.* 2001) have been proposed, as by VanRaden (2008), or, in a crossbred context, by Harris and Johnson (2010). Here we suggest modifications to account for population structure and to rescale these NRM and we compare them with the pedigree NRM **A** (Henderson 1976) in a NZ sheep data set. Getting NRM closely resembling **A** may be desirable because, in a crossbred or multi-breed context, the elements of **A** (particularly off-diagonal elements between breeds) are often closer to the true average identity by descent between individuals. Nevertheless increased similarity with **A** does not equal increased accuracy in GEP.

MATERIALS AND METHODS

Data. A data set of $n = 7{,}855$ animals, sourced from industry and research flocks, with Sheep Improvement Ltd (SIL, http://www.sil.co.nz) pedigree records and Illumina OvineSNP50 BeadChip (http://www.illumina.com) genotypes was used for this analysis. The animals were mostly sires born between 1986 and 2010 of pure and composite recorded breeds of Romney (R), Coopworth (C), Perendale (P) and Texel (T).

Genotypes were cleaned (Dodds *et al.* 2009), which included filtering SNP on call rate, quality score (from the Illumina scoring algorithm), monomorphism, and extreme departure from Hardy-Weinberg Equilibrium (HWE). Any SNP that were not retained as part of the ovine HapMap project (http://www.sheephapmap.org) or were denoted or appeared to be X-linked (including pseudo-autosomal) were removed, leaving $k = 47{,}084$ SNP of the initial 53,903 SNP on the chip.

The pedigree was extracted as deep as possible (up to 23 generations) from the SIL database and consisted of 41,087 animals (including the 7,855 genotyped animals) born between 1969 and 2010. No effort was made to correct the recorded pedigree using the SNP genotypes.

Breed groups. Animals were assigned to 6 groups according to the following definitions. ‘Pure’ bred R, C, P and T were defined as being ≥ 75% of that breed. Two groups of composites were defined for those animals not achieving this purity definition. cRCP have ≥ 50% of R, C and P combined, and < 25% T. cRCPT have ≥ 50% of R, C and P combined, and ≥ 25% T. These definitions were applied after decomposing recently developed ‘breeds’ into their R, C, P and T components, as far as possible, by estimating their breed proportions using the methodology presented in Dodds *et al.* (2012). Table 1 shows the number of animals per breed group.

Table 1. Number of animals per breed group

Breed group	Total	R	C	P	cRCP	T	cRCPT
Number of animals	7,855	4,270	1,697	551	777	317	243

Statistical tools and notation. Here the sum of the elements of matrix $\mathbf{X}$ (or vector $\mathbf{x}$) is denoted $\sum x_{ij}$ (or $\sum x_i$), omitting index and bounds of summation when appropriate. The mean of the elements of $\mathbf{X}$ (and similarly for vectors) is denoted $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}} = \sum x_{ij}/N$, where N is the number of elements of $\mathbf{X}$. The difference between NRM was assessed using the mean absolute difference (MAD), calculated as $MAD(\mathbf{X},\mathbf{Y}) = \sum\left|x_{ij} - y_{ij}\right|/N$ for matrices $\mathbf{X}$ and $\mathbf{Y}$ (and likewise for 2 vectors). Data manipulation and analysis was done in R (R Development Core Team 2012).

Measures of relatedness between individuals. A was computed for all 41,087 animals and only the sub-matrix corresponding to the 7,855 animals genotyped was kept.

Genomic NRM ($\mathbf{G}$) were obtained using the methods described below. Care was taken to use only methods producing (in the worst case semi) positive definite $\mathbf{G}$. First, from VanRaden (2008):

$$\mathbf{G}_a = \mathbf{Z}_a\mathbf{Z}_a'/(2\sum p_j(1-p_j))$$

where $\mathbf{Z}_a = \mathbf{M} - 2\mathbf{P}$, with $\mathbf{M}$ being the $n \times k$ matrix of SNP genotypes m_{ij} scored as 0, 1 or 2 for animal *i* with respectively a BB, AB or AA call for SNP *j*, and $\mathbf{P}$ a matrix of allele frequencies (AF), whose column $\boldsymbol{p} = \mathbf{1}p$, with $\mathbf{1}$ a vector of 1’s of size *n* and *p* the frequency of the ‘A’ allele for a SNP, calculated on the entire population. A second matrix $\mathbf{G}_b$ was created as:

$$\mathbf{G}_b = \mathbf{Z}_b^*\mathbf{Z}_b^{*\prime}$$

where $\mathbf{Z}_\mathrm{b} = \mathbf{M} - 2\mathbf{P}_\mathrm{b}$ and $\mathbf{P}_\mathrm{b} = \boldsymbol{\Lambda}\mathbf{P}_{\boldsymbol{\lambda}}$ is a $n \times k$ matrix of AF pertaining to each animal, $\boldsymbol{\Lambda}$ being a $n \times l$ matrix whose element λ_{ij} is the proportion of breed *j* for animal *i* for a total of *l* breeds (fixed *a priori*), and $\mathbf{P}_{\boldsymbol{\lambda}}$ a matrix of AF estimates for each SNP and each breed. $\mathbf{Z}_\mathrm{b}^*$ is a rescaled version of $\mathbf{Z}_\mathrm{b}$ so that each element $\mathrm{z}_{\mathrm{b}_{\mathrm{ij}}}^* = \mathrm{z}_{\mathrm{b}_{\mathrm{ij}}}\Big/\sqrt{\mathbf{2}\sum_{j=1}^{k} p_{b_{ij}}\left(1 - p_{b_{ij}}\right)}$, with $p_{b_{ij}}$ being the element of $\mathbf{P}_\mathrm{b}$ relating to animal *i* and SNP *j*. This method of calculation tries to account for AF differences between breeds when estimating $\mathbf{G}$ in multi-breed populations. This topic has been discussed extensively in Harris and Johnson (2010). Next, 2 variations of the 2 methods above were devised where we tried to rescale $\mathbf{G}$ so that $\bar{\mathbf{G}} = \bar{\mathbf{A}}$. The first variation is a convex combination of $\mathbf{G}$ and a constant β:

$$\mathbf{G}_x^* = \pi\mathbf{G}_x + (1-\pi)\mathbf{1}\mathbf{1}'\beta$$

where $x = a$ or b, $\pi \in [0,1[$ and $\beta = (\bar{\mathbf{A}} - \pi\bar{\mathbf{G}})/(1-\pi)$. The second variation rescales $\mathbf{G}$ by using adjusted AF. The expected contribution $E[c_{ijk}]$ of 1 SNP k in HWE to element g_{ij} of $\mathbf{G}$ relating to animals *i* and *j* is 0 if they are unrelated and come from the same population. This can be tested by noting that $E\left[c_{ijk}\right] = E[\mathbf{vv}'] = \sum(\mathbf{uu}' \circ \mathbf{vv}') = 0$, where $\mathbf{u} = (1-p^2, 2p(1-p), p^2)$ the vector of genotype probabilities for a bi-allelic marker under HWE, $\mathbf{v} = (-2p, 1-2p, 2-2p)$ the vector of centred genotypes and the operator ○ denotes the Hadamard (entrywise) product. If we adjust the AF used when calculating $\mathbf{G}$ by adding a constant δ, the expected contribution of 1 SNP k in HWE is now $E\left[c_{\delta_{ijk}}\right] = \sum(\mathbf{uu}' \circ \mathbf{v}_\delta\mathbf{v}_\delta') = 4\delta^2$, where $\mathbf{v}_\delta = (-2(p+\delta), 1-2(p+\delta), 2-2(p+\delta))$. For $\mathbf{G}_a$, we can now choose δ_a as a root of the quadratic equation:

$$2k\delta_a^2 - (\bar{\mathbf{A}} - \bar{\mathbf{G}}_a)\sum_{j=1}^{k}(p_j + \delta_a)\big(1 - p_j - \delta_a\big) = 0$$

to construct $\mathbf{G}_a^{**}$ as $\mathbf{G}_a$, but replacing $\mathbf{P}$ by $\mathbf{P}^* = \mathbf{P} + \mathbf{11}'\delta_a$. Similarly for $\mathbf{G}_b$, δ_b satisfies equation:

$$2k\delta_b^2 - (\bar{\mathbf{A}} - \bar{\mathbf{G}}_a)\Sigma = 0$$

where $\Sigma = (\sum_{i=1}^{n}\sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{k}} 2(p_{b_{ij}} + \delta_b)\left(1 - p_{b_{ij}} - \delta_b\right))/n$ with $p_{b_{ij}}$ the element of $\mathbf{P}_\mathrm{b}$ relating to animal *i* and SNP *j*, and $\mathbf{G}_b^{**}$ can be constructed as $\mathbf{G}_b$ but replacing $\mathbf{P}_b$ by $\mathbf{P}_b^* = \mathbf{P}_b + \mathbf{11}'\delta_b$.

RESULTS AND DISCUSSION

Mean pedigree relatedness between groups. Overall, $\bar{\mathbf{A}} = 0.0048$. Table 2 reports the mean inbreeding coefficient ($\bar{\mathbf{f}} = \overline{d\imath ag(\mathbf{A}) - \mathbf{1}}$) and $\bar{\mathbf{A}}$ within (ignoring the diagonal) and between breed groups. Relatedness within groups ranged from 0.008 (R) to 0.028 (C and cRCPT). Relatedness between groups ranged from < 5 × 10^{-4} to 0.015 (C × cRCP and C × cRCPT).

Table 2. $\bar{\mathrm{A}}$ (%) and $\bar{\mathrm{f}}$ (%) overall between breed groups

Breed group	$\bar{\mathbf{f}}$	R	C	P	cRCP	T	cRCPT	Total
R	2.0	0.8	0.0	0.0	0.3	0.0	0.1	
C	2.4		2.8	0.0	1.5	0.1	1.5	
P	1.1			1.6	0.1	0.0	0.0	
cRCP	1.1				1.5	0.2	1.2	
T	1.2					2.2	1.0	
cRCPT	0.9						2.8	
Total	1.8							0.5

Mean genomic relatedness and comparison with pedigree relatedness. Overall, $\bar{\mathbf{G}}_a = 0.0000$ and $\bar{\mathbf{G}}_b = 0.0018$. The roots δ_a for the AF adjustment were $\delta_a = (-0.0211, 0.0212)$. The roots δ_b were $\delta_b = (-0.0166, 0.0166)$. Fixing $\pi = 0.99$ so that the diagonal of $\mathbf{G}_x^*$ are not shrunk down excessively, using any of the 2 methods of rescaling $\mathbf{G}$ and any estimate of δ_x lead to virtually the same matrix, as $MAD(\mathbf{G}_x^*, \mathbf{G}_x^{**})$ ranged from 3.6 × 10^{-3} to 8.4 × 10^{-3}. It is worth noting that $\pi = 0.99$ is not the value of π minimising $MAD(\mathbf{G}_x^*, \mathbf{A})$. These are $\pi_a = 0.1106$ and $\pi_b = 0.2419$, that produce $\mathbf{G}_a^*$ and $\mathbf{G}_b^*$ that are unreasonably shrunk (ideally, diagonal elements should be kept ≥ 1 as much as possible), because $\mathbf{A}$ is very sparse. A potential improvement would be to minimise $MAD(\mathbf{G}_x^*, \mathbf{A})$ only for elements of $\mathbf{A}$ reaching a certain threshold. Table 3 reports $\bar{\mathbf{f}}$, $\bar{\mathbf{G}}$ and $MAD(\mathbf{G}, \mathbf{A})$ within and between groups respectively for $\mathbf{G}_a$, $\mathbf{G}_b$, $\mathbf{G}_a^*$ and $\mathbf{G}_b^*$. $MAD(\mathbf{G}_a, \mathbf{A})$ within Texel (24.6%) and between groups with Texel or cRCPT was very high. Using $\mathbf{G}_a^*$ instead of $\mathbf{G}_a$ slightly increased $MAD(\mathbf{G}, \mathbf{A})$ within breed (up to 0.4%), but somewhat reduced $MAD(\mathbf{G}, \mathbf{A})$ overall (0.1%) and between breeds (up to -0.5%). Using $\mathbf{G}_b$ reduced overall and per breed $MAD(\mathbf{G}, \mathbf{A})$ and $MAD(\mathbf{f}_\mathbf{G}, \mathbf{f}_\mathbf{A})$ dramatically compared to $\mathbf{G}_a$, especially for Texel (2.9%). $\mathbf{G}_b^*$ lead to a slight decrease in $MAD(\mathbf{G}, \mathbf{A})$ over $\mathbf{G}_b$ (0.1%). The values of $MAD(\mathbf{f}_\mathbf{G}, \mathbf{f}_\mathbf{A})$ and $MAD(\mathbf{G}, \mathbf{A})$ within breed obtained with $\mathbf{G}_a$ (and $\mathbf{G}_a^*$) were very highly correlated with $MAD(\mathbf{p}_{\mathrm{total}}, \mathbf{p}_{\mathrm{breed}})$, the MAD between AF calculated overall and per breed, with correlations of respectively 0.935 and 0.958. Together with the extremely high $MAD(\mathbf{G}_a, \mathbf{A})$ in Texel, this suggested that $\mathbf{G}_a$ (and hence $\mathbf{G}_a^*$) is not well suited to predict $\mathbf{A}$ in a crossbred situation. $\mathbf{G}_b$ and $\mathbf{G}_b^*$ on the other hand predicted $\mathbf{A}$ reasonably well. The impact of rescaling the matrices was marginal.

Table 3. Within breed group $\bar{\mathbf{f}}$, $MAD(\mathbf{f}_\mathbf{G},\mathbf{f}_\mathbf{A})$, $\bar{\mathbf{G}}$ and $MAD(\mathbf{G},\mathbf{A})$, and between group $\bar{\mathbf{G}}$ (above diagonal) and $MAD(\mathbf{G},\mathbf{A})$ (below diagonal) using different G, all in %

G	Breed group	$\bar{\mathbf{f}}$	$MAD(\mathbf{f}_\mathbf{G},\mathbf{f}_\mathbf{A})$	$\bar{\mathbf{G}}$	$MAD(\mathbf{G},\mathbf{A})$	R	C	P	cRCP	T	cRCPT
$\mathbf{G}_a$	R	1.5	2.4	2.7	2.4		-3.2	-0.7	-1.7	-4.8	-3.3
	C	4.7	2.9	6.7	4.1	3.3		-1.4	2.4	0.5	3.0
	P	6.9	6.1	8.5	7.0	1.1	1.5		0.0	-0.2	-0.7
	cRCP	2.3	2.8	1.6	1.6	2.4	2.1	1.4		1.6	1.9
	T	21.6	20.4	26.7	24.6	5.1	1.6	1.0	2.8		9.9
	cRCPT	4.0	3.3	6.3	3.7	3.5	2.3	1.3	1.8	8.9	
	Total	4.5	4.4	0.0	2.8						
$\mathbf{G}_a^*$	R	1.0	2.5	3.2	2.7		-2.7	-0.2	-1.2	-4.3	-2.7
	C	4.1	2.6	7.1	4.5	2.8		-0.9	2.8	1.0	3.4
	P	6.3	5.5	8.9	7.4	1.0	1.2		0.5	0.3	-0.2
	cRCP	1.8	2.6	2.1	1.6	2.1	2.2	1.4		2.1	2.4
	T	20.9	19.7	27.0	24.8	4.6	1.7	0.9	2.9		10.3
	cRCPT	3.4	2.9	6.7	4.1	3.0	2.5	1.1	1.9	9.3	
	Total	4.0	4.3	0.5	2.7						
$\mathbf{G}_b$	R	3.1	3.1	0.0	2.0		0.0	0.0	0.1	0.0	0.0
	C	2.1	2.6	1.4	2.2	0.8		0.0	1.2	-0.2	1.1
	P	-0.1	3.1	-0.2	2.6	0.8	0.7		0.0	0.0	0.0
	cRCP	1.4	2.4	1.6	1.3	1.0	1.3	0.8		-0.4	1.1
	T	0.1	3.6	1.8	2.9	0.7	0.8	0.6	1.2		0.2
	cRCPT	-1.3	3.1	2.3	1.5	0.8	1.3	0.7	1.2	1.6	
	Total	2.6	3.3	0.2	1.2						
$\mathbf{G}_b^*$	R	2.3	3.0	0.3	1.9		0.3	0.3	0.4	0.3	0.3
	C	1.4	2.8	1.7	2.1	0.8		0.3	1.5	0.2	1.4
	P	-0.8	3.3	0.1	2.4	0.8	0.7		0.3	0.3	0.3
	cRCP	0.6	2.5	1.9	1.3	1.0	1.3	0.8		-0.1	1.3
	T	-0.6	3.9	2.1	2.8	0.7	0.8	0.6	1.1		0.5
	cRCPT	-2.0	3.6	2.5	1.4	0.8	1.3	0.7	1.1	1.5	
	Total	1.9	3.3	0.5	1.2						

ACKNOWLEDGMENTS

Financial assistance was provided by Ovita Ltd (consortium between AgResearch Ltd and Beef + Lamb NZ). Data was made available by SIL.

REFERENCES

Dodds K. G., Auvray B., Pickering N. and McEwan J. C. (2009) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18**: 296.

Dodds K. G., Auvray B., Newman S.-A. N. and McEwan J. C. (2012) *Australasian Applied Statistics Conference (GenStat & ASReml), Queenstown, New Zealand.* http://www.aasc2012.com/file/Dodds-2012-AASC-talk.pdf. (Accessed 24 January 2013.)

Harris B. L. and Johnson D. L. (2010) *J. Dairy Sci.* **93**: 1243.

Henderson C. R. (1976) *Biometrics* **32**: 69.

Meuwissen T. H., Hayes B. J. and Goddard M. E. (2001) *Genetics* **157**: 1819.

R Development Core Team (2012) 'R: A language and environment for statistical computing' R Foundation for Statistical Computing, Vienna, Austria.

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