New Method to Combine Pedigree and Molecular Relationships Applied to Brandrood Cattle

E. Bömcke[*,†], J. J. Windig[‡] and N. Gengler[*,§]

Introduction

Relationship coefficients correspond to genetic covariances between related individuals expressed relatively, independently from the considered traits. Relationship coefficients are traditionally based on pedigree data. But any pedigree is somewhat incomplete for various reasons. The two most important are, first, that cut-off dates for recording the pedigree might exist and, second, that animals of unknown origins enter the herdbooks, e.g., animals coming from another breed/country. With availability of molecular data, pedigree based relationship coefficients are often completely replaced by molecular coefficients. Examples are relationships from microsatellites for bio-diversity studies (Caballero and Toro (2002); Oliehoek et al. (2006)) but also genomic relationships from single nucleotide polymorphisms (SNP) as currently used in genomic prediction of breeding values (Zhang et al., 2007; VanRaden, 2008). The availability of dense molecular markers like SNP has opened the door for genomic selection (Meuwissen et al. (2001)). Even if this new evaluation method is very promising for animal breeding, the limit is that genotyping an entire population is impossible due to its high cost or for logistic restrictions (i.e., culled, slaughtered or foreign animals). The current used approach consists in genotyping precisely evaluated animals to create a training population, often of older sires. The developed prediction equations are then used to evaluate prospective future selection candidates. However, this way to proceed is suboptimal as molecular data is not directly combined with pedigree and phenotypic data (e.g., Legarra et al. (2009)). Already VanRaden in 2008 showed equivalences between models using linear prediction equations and models using directly a genomic relationship matrix. Misztal et al. (2009) and Aguilar et al. (2010) developed this further to show that by modifying mixed model equations replacing the relationship matrix A by a modified one that includes genomic information, will lead to genomic predictions including molecular, pedigree and phenotypic data. In many situations, optimal combination of pedigree and molecular data would be the best solutions. This is the case of the Deep Red cattle, a breed originated from the Netherlands, which is incompletely genotyped and has sparse pedigree information. The objective of this study was therefore to apply a new method to estimate relationship by combining molecular data with pedigree data. The estimation of relationships in the Deep Red Cattle population could help for the management of the breed and its conservation.

[*] Animal Science Unit, Gembloux Agro-Bio Tech, University of Liege, 5030 Gembloux, Belgium

[†] Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture (FRIA), 1000 Brussels, Belgium

[‡] Animal Breeding and Genomics Centre, Animal Sciences Group, Wageningen University and Research Centre, 8200 AB Lelystad, The Netherlands

[§] National Fund for Scientific Research (FNRS), 1000 Brussels, Belgium

Material and methods

Pedigree data. The Deep Red Cattle pedigree includes 6477 animals, 2036 males and 4441 females born between 1945 and 2008. Animals born between 2002 and 2007 where considered as reference population (living and reproducing animals). The deepness of pedigree information was characterized by the number of generation-equivalents (GEQ). This parameter is considered as the best criterion to characterize the quality of the pedigree information (Maignel et al. (1996); Baumung and Sölkner (2003)). The GEQ was computed for each animal as the sum of $(1/2)^n$, where n is the number of generations separating the individual from each known ancestor (Huby et al. (2003)). The inbreeding level of the living population was also calculated. The individual inbreeding coefficient (F) is defined as the probability that an individual has two genes identical by descent (Wright (1931)).

Genotypes. DNA samples were collected from 195 animals. The samples were tested for genetic variation at 16 loci with microsatellite markers: BM1824, BM2113, ETH10, ETH225, ETH3, INRA23, SPS115, TGLA22, TGLA126, TGLA227, TGLA53, BM1818, CSRM60, CSSM66, HAUT27, ILSTS006. All these markers are used globally for routine bovine genotyping for various purposes such as parentage verification and kinship analysis (van de Goor et al. (2009)). 52 animals were genotyped only for the first 11 markers. For each marker, the average polymorphism information content (PIC) value was calculated. This parameter was introduced by Botstein et al. (1980) and refers to the value of marker informativeness within a population, depending on the number and allele frequencies.

Methods. Pedigree and markers were combined using the method presented by Bömcke et al. (2009). The used formula is the following:

$$\hat{a}_{xy,combined} = \hat{b}_0 + \sum_{l=1}^{nl} (\hat{b}_l \times ta_{xy,l}) + \hat{b}_p \times \hat{a}_{xy,P}$$

where $\hat{a}_{xy,combined}$ is the estimated combined relationship coefficient between individual x and individual y. b_0 is the estimated intercept. b_l are the nl estimated marker regression coefficients, one for each marker. $ta_{xy,l}$ is the total allelic relationship between x and y. b_P is the estimated pedigree regression coefficient. And $a_{xy,P}$ is the pedigree relationship coefficient. The markers regression coefficients were estimated thanks to the PIC value of the marker and to the number of genotyped marker. The intercept was calculated with the inbreeding level of reference population. Finally, the pedigree regression coefficient was estimated according to the mean GEQ value. As described, the procedure will only combine relationship coefficients of genotyped animals. However, the relationships between ungenotyped and genotyped animals as well as among related ungenotyped animals could be affected by the modification of relationships among genotyped animals. Modification of the complete A matrix is therefore required, the resulting relationship matrix will be called "modified" relationship matrix ($\mathbf{A_{modified}}$). The following equation gave the inverted complete $\mathbf{A_{modified}}$ matrix (Aguilar et al. (2010), Bömcke et al. (2009)):

$$\left(\mathbf{A}_{modified}\right)^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{A}_{combined}\right)^{-1} - \left(\mathbf{A}_{genotyped}\right)^{-1} \end{pmatrix}$$

where the elements of $(\mathbf{A_{combined}})^{-1}$ are obtained through inversion of the combined matrix and $\mathbf{A_{genotyped}}$ is the pedigree based relationship matrix among genotyped animals.

Statistical analyses. Ancestors of animals with the deepest pedigree were removed randomly in order to create a more incomplete pedigree than the real one. The method was tested for its capacity to recreate a pedigree that is at least as good as the real one. In order to test this new method for combining relationships, we calculated the relative errors between "true" pedigree-based relationship values, and estimations obtained with the presented method. Calculation of relative error was based on the Frobenius norm, as described by Misztal et al. (1995). This method is convenient for comparison of covariance matrices with similar diagonal elements. The biases of both estimations as well as the probability density function of the residuals were also calculated.

Results and discussion

The inbreeding level of the reference population was 0.01 with a maximum of 0.32. The mean GEQ of the reference population is equal to 2.82. PIC values of the markers ranged from 0.007 to 0.0479.

Consequently, the estimated intercept was equal to 0.129 for the two sets of markers. The estimated regression coefficients varied with the number of markers genotyped for the compared individuals (11 or 16 microsatellites). The marker regression coefficients ranged from 0.008 to 0.056 when 16 markers were used, while they ranged from 0.009 to 0.061 when only 11 markers were used. The pedigree regression coefficient was equal to 0.378.

After addition of the combined information in the complete relationship matrix, the mean relative error between pedigree relationship values, calculated on complete Deep Red cattle pedigree, and estimations obtained with the incomplete pedigree is equal to 0.902 while the mean relative error between pedigree relationship coefficients and estimations obtained in $A_{modified}$ is equal to 0.607. These values quantified the dispersion of the results, there is thus a strong decrease of the dispersion of the results when we combined pedigree and genotypes in the same matrix. The total mean bias of estimations calculated with classical tabular method on the incomplete pedigree was equal to 0.243 while bias of estimations obtained with modified matrix was equal to 0.149. Again, there is a strong decrease between the two values. These observations are reflected in figure 1.
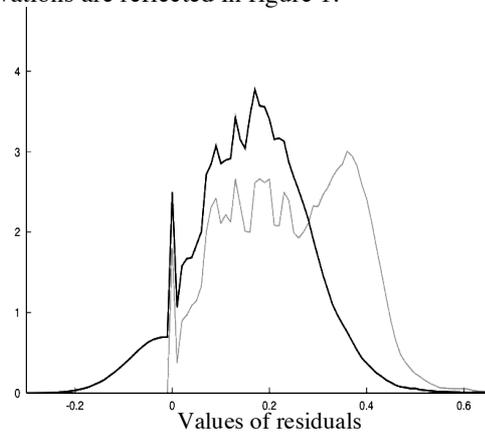
Figure 1: Probability density function of the residuals obtained with modified matrix (in black) and with classical tabular method (in grey) applied on incomplete pedigree.

With the incomplete pedigree, the estimated values are always inferior, i.e. the curve starts at 0, to the expected values, due to the deletion of link in the pedigree, what creates a high dispersion of the results and a high bias. The apparition of residuals inferior to 0 with modified estimations expressed that including the genotypes allows to increase the knowledge of relationship in the population. Another reason is that relationships can be by chance not only lower but also higher than the theoretical value.

The method was applied on the complete pedigree, the inclusion of marker information did not increase significantly the inbreeding level of the population. If the increase had been significant, it would have been necessary to adopt an iterative solution, i.e. re-estimate the intercept and recalculate the combined matrix until there is no significant change in the inbreeding level.

Conclusion

These results clearly showed that there is an interest to combine pedigree and marker information when both sources of information are available. Combining marker and pedigree information produced more accurate A-matrices. This method would be particularly useful for the genetic management of small cattle breed like the Deep Red Cattle but can be extend to other species.

References

Aguilar, I., Misztal, I., Johnson, D. L. *et al.* (2010). *J. Dairy Sci.*, 93:743-752.

Baumung, R. and Sölkner, J. (2003). *Genet. Sel. Evol.,* 35:369-383.

Bömcke, E., Soyeurt, H., Szydlowski, M. *et al.* (2009). *60th Meeting of European Association for Animal Production. Barcelona, Spain*.

Caballero, A. and Toro, M. A. (2002). *Conserv. Genet.,* 3:289-299.

Goor van de, L.H.P., Panneman, H., and Haeringen van, W.A. (2009). *Anim. Genet.,* 40:630-636.

Huby, M., Griffon, L., Moureaux, S. *et al*. (2003). *Genet. Sel. Evol.*, 35:637-655.

Legarra, A., Aguilar, I. and Misztal, I. (2009). *J. Dairy Sci.*, 92:4656-4663.

Maignel, L., Boichard, D. and Verrier, E. (1996). *Interbull Bull.*, 14:49-54.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001). *Genetics*, 157:1819-1829.

Misztal, I., Weigel, K., and Lawlor T.J. (1995). *J. Dairy Sci.*, 78:1862-1872.

Misztal, I., Legarra, A. and Aguilar, I. (2009). *J. Dairy Sci.*, 92:4648-4655.

Oliehoek, P. A., Windig, J. J., van Arendonk, J. A. M. *et al.* 2006. *Genetics*, 173:483-496.

VanRaden, P. (2008). *J. Dairy Sci.*, 91:4414-4423.

Wright, S. (1931). *Genetics*, 16:97-159.

Zhang, Z., Todhunter, R. J., Buckler, E. S. *et al*. (2007). *J. Anim. Sci.*, 85:881-885.