Genomic Selection in Admixed Populations

***R. Rekaya*[1], E. Hay**[1], **S. E. Aggrey**[2]
[1]Department of Animal and Dairy Science, University of Georgia, Athens, Georgia, USA, [2]NutriGenomics Laboratory, Department of Poultry Science, University of Georgia, Athens, Georgia, USA

ABSTRACT: Genomic wide evaluation methods are often conducted using purebred populations. Training and validation are carried out using a select set of pure bred animals. This approach fails when these SNP estimates are used for genomic prediction in other breeds. In this study, we proposed a multi-compartment model where the effect of an SNP could be different between breeds. A simulation was carried out using an admixed population of two divergent lines (A and B) genotyped for 300 markers. Divergence between the two lines was artificially created by multiplying marker effects in one line by a variable α which was sampled from different uniform or normal distributions. The proposed method was compared to the pooled data approach based on the accuracy of predicting the true breeding values. The prediction accuracy using the pooled data approach for line A, was 0.57, 0.49 and 0.30 when α was generated from a uniform distribution between [-2, 2], [-4, 4] and α~ [-8, 8] respectively. Using our proposed method, the corresponding accuracies were 0.65, 0.62 and 0.57, respectively. Similar trend was observed for line B.

Keywords: Genomic selection; Admixed population; SNP

Introduction

Genomic selection benefit is a more accurate pre-selection of animals that inherited genes or chromosome segments of superior merit (Meuwissen et al. (2001)). In livestock, genomic selection is becoming a routine technique, mainly due to the decreasing cost of genotyping for large number of single nucleotide polymorphism (SNP) markers. Genomic selection uses markers that are in linkage disequilibrium with QTLs to estimate breeding values. Currently, genomic selection is conducted on purebreds, and training and validating on such data is successful. However, when training on purebreds and validating on admixed or crossbred animals, this method fails at different degrees depending on the genetic similarity between breeds in the mixture. The main reason genomic selection is not as successful when predicting genetic merit of admixed or crossbred animals is the change in linkage disequilibrium (LD), linkage phase, and allele frequencies between breeds (Roos et al. (2008)).

Accuracy of genome wide evaluation methods crucially depends on the extent of LD between markers and QTLs (De Roos 2009) as well as the size of the reference population. Availability of large reference population is not always guaranteed especially for breeds with limited number of genotyped and phenotyped animals (VanRaden et al. (2009); Hayes et al. (2009)). A plausible solution is the pooled data approach where multiple breed data is pooled to create a large reference population. Intrinsically, the pooled data method assumes constant SNP effects across breeds. This assumption is seldom true due to changes in several parameters such as minor allele frequency, strength of LD between markers and QTLs and linkage phases across sub-populations. Several simulation and real data studies have been conducted to evaluate the accuracy of using different training population pooling strategies (Toosi et al. (2010), Daetwyler et al. (2012), Olsen et al. (2012)). Their results showed an increase in prediction accuracy when subpopulations are genetically close. More recently, (Kachman et al. (2013)) showed that using a multi-breed training population did not increase prediction accuracies than single breed training population when the breed had a reasonable number of animals. However the prediction accuracy increased for breeds with small number of genotyped animals. Similar results have been reported in dairy cattle (Pryce et al. (2011)). When the number of genotyped animals is large enough, within breed training and validating results in the highest accuracy. Furthermore, accuracies dropped when breeds other than the one used in the validation were included in the training set (Toosi et al. (2010), Hayes et al. (2009), Ibanez-Escriche et al. (2009)).

The objective of this study is to develop a model that benefits from the increased population size of the pooling approach and allows for breed specific SNP effects. This will be accomplished through the re-parameterization of SNPs in the admixed population.

Materials and Methods

Statistical method

Development of a multi-compartmental model for genomic selection in admixed populations:

SNP effects change across breeds or crossbred groups due to several factors including, change of minor allele frequency, strength of LD between markers and QTLs, and linkage phase between marker and QTL alleles. We will refer to breeds or crossbred groups in an admixed population as lines. Our idea is based on the possibility of inferring change in SNP effects between lines as a function of their genetic similarity. Our hypothesis is that the genetic similarity between two lines could be either directly modeled through a one–to-one mapping function between SNP effects or indirectly by using information already available in the SNP genotype data.

General idea: Without loss of generality, assuming an admixture population of two lines, the pooled data approach postulates that the SNP effects to be constant across lines which is not true and depends on the genetic similarity between the lines. Using such approach, the combined data will be analyzed using the following model:

$$y_{ij} = \mu + \sum_{k=1}^{p} X_{ik} g_k + e_{ij} \qquad [1]$$

where y_{ij} is the phenotype (or pseudo-phenotype) for animal *i* in line j (j=1,2), μ is the overall mean, X_{ik} is the genotype for animal *i* at locus k (k=1,2,..,p), g_k is the k^{th} SNP effect, and e_{ij} is the residual term. A more realistic model will be to assume different SNP effects between the two lines leading to:

$$y_{i1} = \mu + \sum_{k=1}^{p} X_{ik} g_k + e_{i1} \quad [2]$$

$$y_{i2} = \mu + \sum_{k=1}^{p} X_{ik} g_k^* + e_{i2} \quad [3]$$

Where y_1= (y_{11}, y_{21},…….., y_{n1}) and y_2= (y_{12}, y_{22},…….y_{n2}) are the vectors of observations for line 1 and line 2, respectively, g_k and g_k* are the k^{th} SNP effects in lines 1 and 2, respectively.

Furthermore, g_k* can be written as a linear function of g_k

$$g_k^* = \alpha_k g_k$$

and equation in [3] becomes

$$y_{i2} = \mu + \sum_{k=1}^{p} X_{ik} (\alpha_k g_k) + e_{i2} \quad [4]$$

Where α_k is an unknown real number indicating the similarity of the effect of SNP_k between the two lines. Consequently, the model in equations [2] and [4] can be rewritten in matrix notation as:

$$y = \mu 1_n + X^* g + e \quad [5]$$

Where **y** is the vector of observations for both lines, **g** is the vector of SNP effects in line 1; $\mathbf{X}^*$ is a modified matrix of SNP genotypes where the elements in rows corresponding to individuals in line 2 are multiplied by their respective α_k. If all α_k are equal to one, the matrix $\mathbf{X}^*$ will be identical to the original matrix of SNP genotypes, **X**, as is the case in the pooled data approach. If the vector **α** is known, the implementation of model in [4] is straightforward using any of the existing methods for genome wide association. Unfortunately, the vector **α** is unknown and the model in [5] is not fully identifiable. Thus, **α** and **g** cannot be uniquely estimated. In order to deal with this non-identifiability of the model, a hierarchical Bayesian approach was adopted.

Simulation

A population of 2,799 animals consisting of two lines A and B was simulated with 1,989 in the first line and 810 animals in the second line. Different line sizes were also tested. The genotypic data consisted of 302 SNPs. SNP effects, g_i, were sampled from a normal distribution $N(0,\sigma_{gi}^2)$. The SNP effects were summed over all SNP genotypes for each animal to compute its true BV. Phenotypes were generated by adding an error term sampled from a normal distribution to each true BV. Heritabilities of simulated traits were set to 0.3 and 0.5

To create the divergence between the two lines, SNP effects of the second line were multiplied by a term α which was sampled either from uniform distributions with bounds set equal to [-2,2], [-4,4], [-8,8], normal distributions; N(1,0.01), and N(1,0.05) or from a mixture of a normal distribution, N(1,0.01), and a degenerative distribution at 1.

Molecular breeding values (MBV) were calculated as:

$$MBV_i = \hat{\mu} + \sum_{j=1}^{p} x_{ij} \hat{\alpha}_j \hat{g}_j$$

Accuracy was computed as the average correlation (over five replicates) between the estimated MBV and the true BV.

Results and Discussion

Correlations between TBV and estimated MBV using pooled data (A+B) for training are presented in Table 1. Using uniform or normal distributions, the correlation decreases with the increase in the variability of α. This is expected because the further α deviate from 1 the smaller the genetic similarity is between the components of the admixed population. In fact, when α was sampled from a uniform [-2, 2], accuracy for line A was 0.65, it decreased to 0.62 when α was sampled from uniform [-4, 4] and then a larger decrease when α was sampled form uniform [-8, 8]. As the two lines diverge, LD profiles are likely to differ or even breakdown. Additionally, LD phases between some markers and QTLs may be reversed across lines (de Roos et al. (2008)). Using our proposed method, there was a substantial increase in accuracies under all simulation scenarios of α as shown in Table 2. In fact, the accuracy increased to 0.65, 0.62 and 0.57 for the three uniform distributions used to generate α, respectively. Similar trend was observed for line B, although the magnitude of the correlations was lower. The same behavior was seen when α was sampled from normal distributions with different variances (Tables 2). Although the increase in accuracy was not as large as when uniform distributions were used, it is still a considerable gain.

Table 1 Correlations between TBV and MBV using Pooled data method.

Distribution of α	Line A	Line B
α ~ [-2,2]	0.45	0.22
α ~ [-4,4]	0.34	0.31
α ~ [-8,8]	0.3	0.21
α ~ N(1,0.01)	0.42	0.27
50% α ~ N(1,0.01)	0.45	0.34
50% α=1		
α ~ N(1,0.05)	0.45	0.32

Tables 3 and 4 present the accuracies between true and estimated breeding values using two heritability values (0.3 and 0.5) and different sample size for line A (1989 or 810 individuals) using the pooled data approach (Table 3) and the proposed method (Table 4) when α was generated either from a uniform [-2, 2] or [-4,4]. For both methods and as expected, accuracies decreased with the decrease in heritability or the population size. However, across all simulation scenarios of heritability and sample size, the proposed method was superior to the pooled data approach. Such superiority ranged approximately from 2% to 20%.

Table 2 Correlations between TBV and MBV using multi-compartment model.

Distribution of α	Line A	Line B
α ~ [-2,2]	0.53	0.35
α ~ [-4,4]	0.53	0.37
α ~ [-8,8]	0.57	0.43
α ~ N(1,0.01)	0.45	0.33
50% α ~ N(1,0.01)	0.46	0.36
50% α=1		
α ~ N(1,0.05)	0.46	0.33

Table 3 Correlations between TBV and MBV using pooled data method.

Pooled method		
h^2 = 0.3		
	Line A= 1989 line B= 810	**Line A= 810 Line B= 810**
α[-2,2]	Line A=0.45	Line A=0.36
	Line B=0.22	Line B=0.23
α[-4,4]	Line A=0.34	Line A=0.25
	Line B=0.31	Line B=0.26

Table 4 Correlations between TBV and MBV using Multi-compartmental model

h^2 = 0.5		
	Line A=1989 line B=810	**Line A=810 Line B=810**
α[-2,2]	Line A=0.58	Line A=0.46
	Line B=0.30	Line B=0.36
α[-4,4]	Line A=0.45	Line A=0.31
	Line B=0.42	Line B=0.41
Multi-compartment model		
h^2 = 0.3		
	Line A= 1989 line B= 810	**Line A= 810 Line B= 810**
α[-2,2]	Line A=0.53	Line A=0.45
	Line B=0.35	Line B=0.34
α[-4,4]	Line A=0.53	Line A=0.41
	Line B=0.37	Line B=0.35
h^2 = 0.5		
	Line A=1989 line B=810	**Line A=810 LineB=810**
α[-2,2]	Line A=0.68	Line A=0.59
	Line B=0.52	Line B=0.52
α[-4,4]	Line A=0.67	Line A=0.56
	Line B=0.53	Line B=0.52

Conclusion

Pooling data from lines or breeds in the training set when conducting genome wide evaluation studies seems an attractive approach since it benefits from the increase in power. Its performance is variable and depends largely on the genetic similarity between the sub-populations in the mixture. The proposed multi-compartment model and based on the simulation results is clearly superior as it allows systematically for the accounting of the difference in SNP effects across lines. Its superiority compared to the pooled data approach ranged from approximately from 2 to 20% and increases as the divergence between lines increases. The current simulation parameters do not reflect the actual SNPs density in commercially used panels. Thus, it is needed that the performance of the proposed model be evaluated when large numbers of SNPs are genotyped.

Literature Cited

Daetwyler, H. D., K. E. Kemper, J. H. J. van der Werf and B. J. Hayes. 2012. *J. Anim. Sci.* 90 (2012), pp. 3375–3384

De Roos, A. P. W., Hayes B. J., Goddard M.E. 2009. *Genetics* 183: 1545-1553

De Roos, A. P. W., Hayes B. J., Spelman R. J., et al. 2008. *Genetics* 179: 1503-1512

Gautier, M., T. Faraut, K. Moazami-Goudarzi, et al., 2007. *Genetics* 177: 1059–1070.

Hayes, B., P. J. Bowman, A.C. Chamberlain, et al. M. E. Goddard. *Genetics Selection Evolution* 2009, 41:51

Ibanez-Escriche, N., R. L. Fernando, A. Toosi, A. et al. *Genetics Selection Evolution* 2009, **41**:12

Kachman S D, M L Spangler, G L Bennett, et al. *Genetics Selection Evolution* 2013, **45**:30

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

Olson KM, VanRaden PM, Tooker ME. *J Dairy Sci* 2012,95:5378-5383

Pryce JE, Gredler B, Bolormaa S. et al. *J Dairy Sci* 2011. 94:2625-2630

Toosi, A., R. L. Fernando, J. C. M. Dekkers. 2010. *J. Anim Sci*. 2010. 88:32-46.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, et al. *J. Dairy Sci.* 2009 2:16-24.