Whole Genome Evaluation For Related Populations

L. Varona[*], C. Moreno[*], N. Ibáñez-Escriche [†], J. Altarriba[*]

Introduction

Availability of dense marker maps of livestock species has opened new possibilities for genetic evaluation of individuals by whole genome association procedures (Meuwissen et al., 2001). However, an accurate prediction of breeding values ussually involves a huge genotyping effort that cannot be easily supported by populations with reduced population size.

In an attempt to improve the accuracy of the genomic prediction for these small populations, De Ross et al. (2009) and Hayes et al. (2009) have proposed combining data set from multiple populations. These authors showed that the predictive ability from the one population to another depends on the genetic distance between populations and the SNP marker density.

The aim of this manuscript is to expand the genomic blup procedure described by Gianola et al. (2003) to a multivariate scope, by considering SNP effects estimated in related population as different but correlated variables.

Material and methods

Simulation. Parameters of simulation were obtained from Meuwissen et al., (2001). Thus, LD was simulated by drift and mutation. First, 1100 generations of random mating were simulated in an effective population 100 (see figure 1). On generation 1000, population 1 was generated by expanded it to 500 males and 500 females that were randomly mated during three generations to obtain a training population of 3000 invididuals. In the case of population 2, three scenarios were considered, following the same process of expansion to 3000 individuals at generations 1005, 1025 and 1100, respectively.

In the simulation, genome consisted of 10 chromosomes of 100 cM each with 1000 loci each one (9000 SNPs and 1000 QTLs in total). Both SNPs and QTLs were biallelic and located at random map positions. Mutations were generated at a rate of $2.5x10^{-3}$ per locus per generation at the markers and at a rate of 2.5×10^{-5} at the QTL. The additive effects were sampled from a standard normal distribution and scaled to get the desired values of h^2 (V_A/V_P) being V_A and V_P the additive and phenotypic variance. In generation 1 about half of the loci were fixed for allele 1 and the other half were fixed for allele 2.

[*] Universidad de Zaragoza, Facultad de Veterinaria, Miguel Servet 177, 50013 Zaragoza, Spain
[†] IRTA. Area de Genètica i Millora, Rovira Roura 191, 25198 Lleida, Spain.

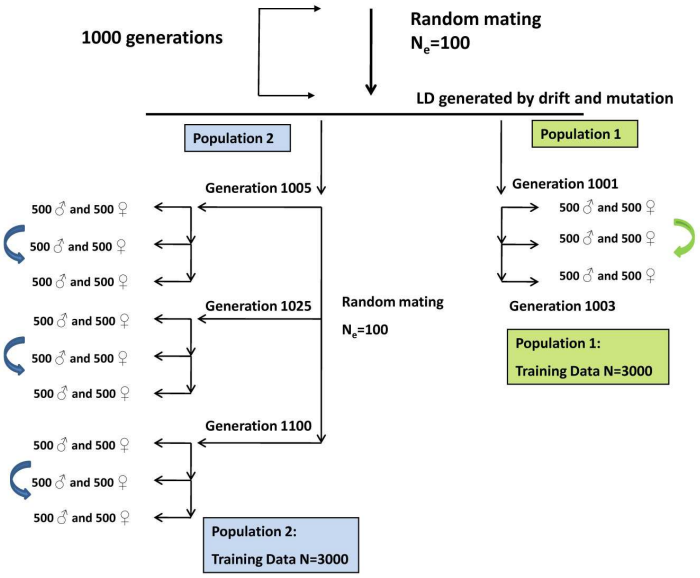


Figure 1. Structure of the simulated population

Statistical analysis. The model of analysis assumed that the phenotypic value of the *ith* individual of the *jth* population is:

$$y_{ij} = \mu_j + \sum_{k=1}^{p} x_{ijk} g_{kj} + e_{ij}$$

Where μ_j is the mean of the *jth* population, p is the number of SNPs, x_{ijk} is a function that indicates the SNP genotypes for the *kth* SNP in the *ith* individual of the *jth* population, g_{kj} is the *kth* SNP effect on the *jth* population and e_{ij} is the *ith* residual on the *jth* population. Prior distribution is assumed uniform between appropriate bounds for population means and Gaussian for the residuals. However, for the SNP effects, three different prior distributions were assumed, determining three different models of analysis.

Model 1: SNP effects from both populations are considered independent

$$\mathbf{g}_1 \sim N\left(0, \mathbf{I}\sigma^2_{g_1}\right) \qquad \mathbf{g}_2 \sim N\left(0, \mathbf{I}\sigma^2_{g_2}\right)$$

Where $\mathbf{g}_1$ and $\mathbf{g}_2$ are the vectors of SNP effects and $\sigma^2_{g_1}$ and $\sigma^2_{g_2}$ are their variances for populations 1 and 2, respectively.

Model 2: The SNP effects for both populations are the same ($\mathbf{g}_1 = \mathbf{g}_2$).

Model 3: The SNP effects are assumed to be different but correlated.

$$\begin{matrix}\mathbf{g}_1 \\ \mathbf{g}_2\end{matrix} \sim N\left(0, \mathbf{I} \otimes \mathbf{G}\right)$$

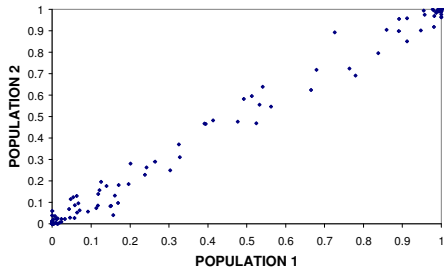
where **G** is the (co) variance matrix for SNP effects.

For each model, the prior distribution of the variance components was assumed flat. The implementation was performed via Gibbs sampling, and posterior estimates were calculated by averaging the samples from 10,000 cycles, after discarding the first 1,000. Each case of simulation was replicated 20 times and the models were compared in terms of accuracy calculated by the correlation between simulated and predicted breeding values.

Results

The results of gene frequencies of the QTL in one replicate with 5 and 100 generations of separation between populations are presented in figure 2. When only 5 generations of a mutation drift process with effective size of 100 separate the populations, gene frequencies of the QTL remained similar, and almost every QTL that segregated in one population also did it on the other. On the contrary, discrepancies in gene frequencies were greater when populations differed in 100 generations and also there was a relevant number of QTL segregating only in one of the populations.

a) 5 generations



b) 100 generations

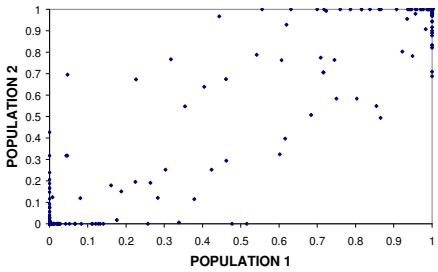


Figure 2. Gene frequencies for the 1000 QTL in populations 1 and 2 when separated by 5 and 100 generations (one replicate) .

The results of accuracy of estimation of breeding values with the three alternative models and the correlation between SNP estimates under Model 3 are presented in table 1. In the cases when the number of generations of separation was low (5 generations), there was almost no differences between Model 2 and Model 3. This fact can be explained because the small differences between QTL frequencies. On the contrary, as the distance between

generations increase, the results of Models 1 and 3 becomes similar, whereas the results of Model 2 were worst.

Table 1: Average (and standard error) of accuracy of GWE with Models 1, 2 and 3 and correlation between SNPs under Model 3.

h^2	G	Model 1	Model 2	Model 3	Corr.
0.15	5	0.685 (0.015)	0.719 (0.014)	0.727 (0.013)	0.777 (0.028)
0.15	25	0.682 (0.013)	0.693 (0.013)	0.703 (0.013)	0.609 (0.049)
0.15	100	0.686 (0.013)	0.653 (0.014)	0.690 (0.013)	0.256 (0.044)
0.35	5	0.773 (0.011)	0.805 (0.010)	0.809 (0.010)	0.851 (0.033)
0.35	25	0.771 (0.011)	0.777 (0.011)	0.792 (0.010)	0.743 (0.038)
0.35	100	0.782 (0.008)	0.756 (0.015)	0.787 (0.008)	0.232 (0.040)

G: Generations of separation

As expected the estimates of the correlation between SNP effects were higher when the populations were separated by few generations and becomed lower as the genetic distance between populations increased.

Conclusion

The results suggest than the information provided by whole genome evaluation on a related population can be used to increase accuracy on a population and reduce genotyping effort. However, misuse of this information can decrease the accuracy of the prediction of breeding values when populations are far related. Moreover, the proposed strategy allows to weight the information provided by a related population. Further research must be done to study the performance of the proposed procedure with different population sizes and markers densities and to explore the possibility of expanding the strategy to more complex methods to predict breeding values from SNP genotypes.

References

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

Gianola, D, Perez-Enciso, M., Toro, M. A. (2003). *Genetics*, 163: 347-365.

Hayes, B. J., Bowman, P. J., Chamberlain, A. C. et al. (2009). *Genet. Sel. Evol.* 41:51

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