Various Statistical Models For Prediction Of SNP Effects From A Multi-National/Breed Reference Population

M.A. Nilforooshan[*], L. Rönnegård[†] and H. Jorjani[*]

Introduction

There has been a major transformation of conventional breeding programs to the breeding programs that make use of genomic information in genetic evaluations. This transformation has been obvious especially in dairy cattle genetics (Dürr, 2009). After a few decades of using markers for QTL mapping and marker assisted selection, now with mapping single nucleotide polymorphisms (SNPs) along the genome and genomic selection, a new era in genetic evaluation of livestock has started.

Genomic selection is based on using dense SNP maps for predicting breeding values of the genotyped animals. The solutions for SNP effects should be generated from animals in the reference population. The reference population is a sample of animals that have been assayed for the markers, recorded for the trait and used for the prediction of genomic effects (Goddard and Hayes, 2009).

Nowadays, the number of parameters whose effect should be estimated is very large. Consequently, the reference population should be large enough to allow these estimations. This limit has been judged to be around 2,500 animals for dairy cattle populations (Schaeffer, 2006). One way to solve this problem is merging populations to create a large reference population, which has an important impact on increasing the accuracies. Huisman et al. (2009) using multiple purebred broiler lines, suggested that a joint reference population can be used to estimate marker effects across lines with high accuracies. In order to increase the accuracy of genomic evaluations of dairy cattle, some countries like the United States and Canada have started joint genomic evaluations (Wiggans et al., 2009) using a shared reference population in the two countries. However, different SNPs or QTLs may react differently in different lines or breeds, and also in across country reference populations due to genotype by environment interaction (G×E). Previous experiences with conventional international genetic evaluations have shown that G×E is an important issue.

The aim of this research was to provide a step by step development of a statistical model for estimation of SNP effects from multi-national/breed reference populations, as well as providing the ability of estimation of SNP effects based on the mean and the scale of different countries in an international genomic evaluation.

[*] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, SE-750 07 Uppsala, Sweden
[†] Department of Statistics, Dalarna University, SE-781 70 Borlänge, Sweden

Methods

M1: Assuming the following simple linear model with no intercept, the effect of each SNP would be predicted using the least square method.

$$y = \Sigma X_j g_j + e$$

where: y is the vector of performance records (phenotypes), X_j is the incidence matrix linking SNPs to animals, g_j is the vector of SNP effects for the j^{th} SNP, and e is vector of random residual effects.

M2: The overall mean of the populations can be included in the model.

$$y = \mu 1_n + \Sigma X_j g_j + e$$

where: μ is the total mean value, and 1_n is the vector of n ones (number of animals in the reference population).

M3: The first correction step for multiple populations is considering the fact that different populations may have different means. Therefore, the SNP effects can be corrected for those means.

$$y_i = B_i \mu_i + \Sigma X_j g_j + e_i$$

where: B_i is the incidence matrix relating animals to the populations, μ_i is the vector of population means for the i^{th} population.

M4: Random polygenic effect of the animal is added to the model.

$$y_i = B_i \mu_i + \Sigma X_j g_j + Zu + e_i$$

where: Z is the incidence matrix linking polygenic effects to animals, and u is vector of polygenic effects

M5: Modification can be made in MME of **M4** for considering different heritabilities for polygenic effects in different populations.

M6: Modification can be made in the equation for considering G×E for polygenic effects.

$$y_i = B_i \mu_i + \Sigma X_j g_j + Z_i u_i + e_i$$

where: u_i is the vector of polygenic effects for the i^{th} population.

M7: Different means for SNP effects in different populations can be added to the equation.

$$y_i = B_i \mu_i + \Sigma X_j g_{j(i)} + Z_i u_i + e_i$$

where: $g_{j(i)}$ is vector of SNP effects for the j^{th} SNP within the i^{th} population.

M8: Starting from **M6**, the equation can be modified to consider random SNP effects (rather than fixed).

$$y_i = B_i \mu_i + Z_i u_i + \Sigma W_j g_j + e_i$$

where: W_j is the incidence matrix linking SNPs to animals

M9: Modification can be made in MME of **M8** for considering different heritabilities for different SNPs in different populations.

M10: Modifications can be made in the equation to include SNP×E interaction

$$y_i = B_i \mu_i + Z_i u_i + \Sigma W_{ij} g_{ij} + e_i$$

where: g_{ij} is vector of the SNP effects for the j^{th} SNP in the i^{th} population.

Discussion

Reference population size is one of the most important factors that influence the reliability of the prediction of SNP effects. The size of some reference populations may not be large enough to provide high accuracies for the prediction of SNP effects. One of the best solutions is merging small reference populations into a large one. However, this practical solution has its own statistical problems. The most important statistical problem is finding a way for considering genotype by environmental interaction among the merged populations.

This research tried to provide a progression of statistical models from simpler to more complicated models that can be used to solve the issues regarding merged reference populations (e.g. different means for different populations, different mean and variances of SNP effects in different populations, different heritabilities and G×E for polygenic effects, and SNP×E interaction).

M1 was the simplest model, including only fixed SNP effects in the model, even without the population mean. In this way, the SNP effects are expressed as deviations from the population mean. Meuwissen et al. (2001) used this method by setting the overall mean to zero. Their explanation was that the overall mean cannot be distinguished exactly from the fixed haplotype effects. Meuwissen and Goddard (1996) also using **M4** did not include the overall mean in the model. In **M2**, a separation was made between SNP effects and the overall mean. In this way, SNP effects were expressed with a mean of zero instead of the overall mean. This method has been used in several QTL studies (e.g., Meuwissen et al., 2001). **M3** took the first step of distinguishing among more than one population. A design matrix (B_i) was used for considering different means for different populations.

M3 was extended to **M4** by including random polygenic effects with no improvement for distinguishing among the populations. **M4** was the first step of upgrading from a least square method to a BLUP or Bayesian method. **M4** has been used in several studies considering one population, using the vector of ones (1_n) instead of B_i design matrix (e.g., Hayes et al., 2009). Calus and Veerkamp (2007) studied the effect of including or ignoring polygenic effects on the accuracy of total breeding values, when there is coverage of the genome with approximately one SNP per cM. They found that polygenic effects can improve the accuracies, especially for low heritable traits and the total genetic variance would be underestimated by ignoring polygenic effects.

From **M4** to **M6** a distinction was made among populations for polygenic effects by considering different heritabilities in different populations (**M5**) and G×E among the populations (**M6**). In **M7**, the first step toward differentiation among SNP effects in different populations was taken by considering different mean of SNP effects in different populations.

Considering SNP effects as fixed is still not a good solution dealing with small populations, because the degree of freedom with which the SNP effects are estimated is low, which will result in low accuracy predictions (Meuwissen et al., 2001). To consider an uncertainty about SNP effects, it is better to consider random distributions for them. In **M8**, an upgrade was

made from **M6** by considering SNP effects as random rather than fixed and considering different heritabilities for different SNPs, but with no distinction for SNP effects among populations. Several studies have adopted both genomic and polygenic effects as random, but for data on one population (e.g., Calus and Veerkamp, 2007; Goddard, 1992; Hayes et al., 2009). **M9** started making a distinction for random SNP effects among different populations, by considering different heritabilities for different SNPs in different populations. The most complete model (**M10**) was built by incorporating SNP×E interaction in **M9**. Therefore, the issues regarding small populations were resolved by merging reference populations into one and considering SNP effects as random, and the statistical issues regarding merged populations were resolved by considering G×E interaction for both polygenic and genomic effects, different heritabilities and different (normal) distributions for SNP effects in different populations.

Conclusion

Merging reference populations is a useful way for improving the accuracy of genomic evaluations, especially for small populations. However, the differences between the genetic background of the merged populations and G×E should be considered.

The possible differences between the results achieved by the series of statistical models would be partly due to the differences among the capabilities and accuracies of the models and partly due to the genetic distances among the merged populations.

The current research could reach to an advanced model which is able to handle data from merged reference populations from different countries or breeds. The developed model, in its current design, can be used for international genomic evaluations, using a unique reference population for estimating SNP effects based on the mean and the scale of different countries.

References

Calus, M. P. L., and Veerkamp, R. F. (2007). *J. Anim. Breed. Genet.*, 124: 362–368.

Dürr, J. W. (2009). In *Proc 60th EAAP*, Book of abstracts No. 15, page 322.

Goddard, M. E. (1992). *Theor. Appl. Genet.*, 83: 878–886.

Goddard, M. E., and Hayes, B. J. (2009). *Nature Reviews Genetics*, 10: 381–391.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J. *et al.* (2009). *J. Dairy Sci.*, 92: 433–443.

Huisman, A. E., Vereijken, A., Van Haandel, B. *et al.* (2009). In *Proc 60th EAAP*, Book of abstracts No. 15, page 211.

Meuwissen, T. H. E., and Goddard, M. E. (1996). *Genet. Sel. Evol.*, 28: 161–176.

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

Schaeffer, L. R. (2006). *J. Anim. Breed. Genet.*, 123: 218–223.

Wiggans, G. R. et al. (2009). *J. Dairy Sci.*, 92: 3431–3436.