SOCIEDADE BRASILEIRA DE MELHORAMENTO ANIMAL – SBMA
IX Simpósio Brasileiro de Melhoramento Animal
20-22 de junho de 2012, João Pessoa, PB, Brasil

9 th Biennial Symposium of the Brazilian Society of Animal Breeding
June 20-22, 2012, João Pessoa, Paraíba, Brazil

FINDING THE BALANCE BETWEEN QUANTITATIVE AND GENOMICS ISSUES – A CASE OF A SINGLE-STEP GENOMIC evaluation (ssGBLUP)

Ignacy Misztal

Department of Animal and Dairy Science, University of Georgia, Athens

Introduction

Most of this introduction is based on my editorial (Misztal, 2007).

Some 20 years ago there was a shift in governmental funding away from animal breeding and quantitative genetics to almost exclusively molecular genetics. The molecular genetics offered the possibility of fast genetic progress based on mainly lab work. Old quantitative work that required extensive training in quantitative genetics, statistics, and programming, seemed obsolete. Great hopes were put into finding markers for major genes (QTL) that could help solve the new challenges. However, few markers/QTLs were detected, that those that were detected had their estimated effects inflated.

The new focus in animal breeding is genomic selection. In this methodology, one estimates effects of individual haplotypes or uses SNP-derived relationships. The genomic selection has an opposite assumption from the previous effort in markers but same as in "black box" genetics: that there are a large number of genes responsible for a trait. The lab work is now done by commercial companies using SNP chips so the old lab skills are mostly obsolete.

The genomic selection brings quantitative genetics bask into focus. The recent studies show that SNP effects are seldom effects of major QTL, that SNP effects are not forever but need to be re-estimated, that large amounts of genomic data is required for accurate predictions, and that simple models can create wildly inflated and inaccurate estimates. Also that poor or antiquated models used in a genetic evaluation result in poor genomic estimates, that GxE is not negligible, and that crossbreeding needs to be accounted for.

This paper focuses on current experiences with the genomic evaluation and especially with the single-step GBLUP. These experiences provide information of which skills are useful for successful animal breeding in the genomic age.

Methods for genomic selection

Currently, the most common scenario in obtaining the genomic prediction via SNP arrays involves a multistep approach. It includes 1) running a regular evaluation, 2) extracting pseudo-observations for genotyped individuals, e.g., daughter deviations or de-regressed EBV, 3) estimating SNP effects using pseudo-observations as records, and 4) possibly combining the genomic predictions with parent average (VanRaden, 2008). For smaller populations, one can run step 3 with phenotypic records; however, the information on relatives is not utilized. Step 3 usually requires estimating weights for SNP effects, mostly via Bayesian procedures such as BayesX (where X could be A, B, C. etc.) (Meuwissen et al., 2001; Hayes et al., 2009). A procedure in which all weights are assumed equal leads to a genomic relationship matrix **G** and is called

GBLUP(Habier et al., 2007).However, when weights are known, methods based on **G** and estimation of SNP effects are equivalent.

Current experiences with SNP panels of around 50-60k indicate that a) at least 1000 genotypes of high accuracy animals are required for a noticeable increase in accuracy and b) GBLUP is almost or as accurate as the BayesX procedures (VanRaden et al., 2009; Hayes et al., 2009). These indicate that the number of QTL in quantitative traits is high (Daetwyler et al., 2010) and subsequently the genomic selection works mostly by capturing relationships than by estimating effects of major genes. Improvements with large SNP of over 500k provide only a small increase of accuracy as the effect of increase SNP can be viewed as reducing a sampling error of **G** to a finite number of SNP. Thus the primary mode of the genomic information in genetic evaluation is improved relationships among animals that also include the information about the Mendelian sampling (Goddard, 2010).

The multi-step methodology is complicated and includes several approximations. Pseudo-observations are dependent on other estimated effects and approximated accuracy of EBV. Models in BayesX are usually simple and single-trait that additionally ignore non-genetic covariances among the residuals. The index is dependent on the approximated accuracy of EBV and is hard to construct. All the approximations reduce accuracy and can inflate GEBV. Also, because of complexity, the multistep approach is prone to errors, which have been observed in many commercial releases in dairy.

As nearly all the genomic information is included in a genomic relationship matrix, Misztal et al. (2009) proposed a single-step methodology where pedigree and genomic relationships are combined into matrix **H**, which is subsequently used in BLUP. Then, compared to a multistep evaluation, step 1 is modified to use matrix **H**, and steps 2-4 are eliminated. Legarra et al. (2009) and Christensen and Lund (2010) developed such a matrix, and Aguilar et al. (2010) demonstrated that a single step methodology can be simple, fast and accurate. The purpose of this paper is to present the theory and experiences with the single step methodology (ssGBLUP).

Single-step genomic evaluation (ssGBLUP)

Legarra et al (2009) developed matrix **H** that combines pedigree and genomic relationships:

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} [\mathbf{G} - \mathbf{A}_{22}] [\mathbf{I} \quad \mathbf{I}] \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & 0 \\ 0 & \mathbf{I} \end{bmatrix}.$$

where subscripts 1 and 2 denoted ungenotyped and genotyped animals, respectively, and **G** is a genomic relationship matrix as in VanRaden (2008). Aguilar et al. (2010) and Christensen and Lund (2010) found that the inverse of matrix **H** as above has a simple form:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}.$$

Replacing $\mathbf{A}^{-1}$ with $\mathbf{H}^{-1}$ in existing software for genetic evaluation or for estimation of variance components makes those programs applicable for genomic studies.

Efficient computation of $\mathbf{H}^{-1}$ requires efficient computation of $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$, where the last matrix is an inverse of a pedigree-based relationship matrix for genotyped animals only. Aguilar et al. (2011b) demonstrated that with efficient computing, such matrices can be created in a few minutes of computer time for about 10k genotypes and in about 1 hr time for 30k genotypes.

Matrix **G** is calculated by matrix multiplication(VanRaden et al., 2008). Let p_j be allele frequency for genotype "2" in marker j, and let m_{ij} be genotypes for i-th animal and j-th marker such that

$$m_{ij} = \begin{cases} 0 - 2p_j & \text{- homozygous 11} \\ 1 - 2p_j & \text{- heterozygous 12 - 21} \\ 2 - 2p_j & \text{- homozygous 22} \end{cases}$$

so that average $m_{i.}$ is 0.Then **G** =**MM**'/k, where the scale parameter *k*is usually computed as:

$$k = 2\sum p_j(1-p_j) .$$

Gene frequencies affect the mean and scale of **G**. When equal gene frequencies are assumed, averages of the diagonal or off-diagonal elements may be much larger than in $\mathbf{A}_{22}$. Scaling **G** by regression on **A**as in VanRaden (2008) may result in **G**not being positive definite. With current allele frequencies, the average off-diagonal elements are close to 0. Matrix **G** can be made compatible with $\mathbf{A}_{22}$ when current allele frequencies are used and when **G** is scaled for same averages of diagonal and off-diagonal elements as $\mathbf{A}_{22}$ (Chen et al., 2011b).

When genotyped animals include clones, **G** as constructed above is singular and cannot be inverted. Therefore, a common strategy is to replace **G** with $\alpha\mathbf{G} + (1-\alpha)\mathbf{A}_{22}$, where α is close to 1.0, e.g., 0.95. In Christensen et al. (2011), the optimal α was 0.8 although variations in accuracy with different α were minimal.

Applications of ssGBLUP

ssGBLUP has been used for several large-scale analyses including dairy (Tsuruta et al., 2011; Aguilar et al., 2011a; VanRaden, 2012), pigs (Forni et al., 2011; Christensen et al., 2012) and chicken (Chen et al., 2011a).Experiences indicated that 1) ssGBLUP is usually as accurate as or more accurate than multistep methods, 2) the inflation of GEBV is usually smaller, and 3) the operation is much simpler. Another reason for using ssGBLUP is ability to account for selection bias when selection is based on genotypes only (VanRaden, 2012).

Some experiences are species related. For example, there are little differences in GEBV with different **G** in dairy, where the "training" animals are high-accuracy bulls. Different **G** affects estimates of variances components, calculated accuracy and to a smaller degree, real accuracy in pigs and chicken (Forni et al., 2011; Simeone et al., 2011). Chen et al (2011b) found that the scale of **G** influences ranking of genotyped versus ungenotyped animals. The optimal **G** would have same averages of diagonals and off-diagonals as $\mathbf{A}_{22}$. Vitezica et al. (2011) derived a formal proof and showed that, under well-formed **G**, ssGBLUP is more accurate and less biased than a multistep approach. The optimal scaling is equivalent to:

Gc= Fst/2+(1-Fst)**G**c

Where Fst is a fixation index of genotyped animals relative to the base population and **G**c is a genomic relationship matrix computed using realized allele frequencies. In practice, Fst is equal to twice the average of off-diagonals of $\mathbf{A}_{22}$.

ssBLUP in chicken

ssGBLUP was applied in chicken for a large experiment (Chen et al, 2011a). However, results of selection over a few generations were short of expectations (Muir et al., 2012). These were caused by many problems including incorrect genotypes, initial methodology issues, use of a large ungenotyped population that had close ties only to the first generations of genotyped animals, and many other reasons (Muir et al., 2012). Therefore, the practical results of ssGBLUP or any other methodology for genomic selection in a commercial situation is dependent on attention to detail,the mature methodology, and knowledge of issues of genomic selection specific to a given population.

Convergence and biases

Poor convergence rate and/or large reranking have been observed in several analyses. These problems were traced to incompatability between **G** and $\mathbf{A}_{22}$. For well formed G and with many generations of complete pedigrees, G and $\mathbf{A}_{22}$ are very similar, with SD < 0.04 (Wang et al., 2011). However, differences exceeding 1.0 are observed in practice. Such differences are due to short or incomplete pedigrees, pedigree mistakes, incorrect assignment of genotypes, poor quality of genotypes, and unaccounted presence of multiple/lines breeds.

In general $\mathbf{A}_{22}$ is affected by the maximum number of generations, the completeness of pedigree. It is unaffected by multiple-breeds or lines. In A, separate lines and breeds are not realted. In general, **G** is unaffected by pedigrees. However, it is affected by gene frequencies, number of SNP, quality of genotypes, and scaling. These factors, when unaccounted for, can cause large differences between **G** and $\mathbf{A}_{22}$, and subsequently poor convergence rate with iterative methods.

One particular source of differences between **G** and $\mathbf{A}_{22}$ and subsequently poor convergence rate is heterogeneous base populations or missing parents at many generations. One remedy here is to cut the number of pedigrees to 4-5 so that the effect of missing parents before is eliminated. Experiences indicate retained or even slightly increased accuracy after cutting.

Another important source of difference is presence of multiple lines or breeds. In such a case, the distribution of the diagonal of **G** may be multi-modal (Simeone et al., 2011a). The solution is to construct **G** reflecting differences among line/breeds, e.g., such as in Harris and Johnson (2010). The differences are different allele frequencies and Fst for each pure line. The issue of multibreed ssGBLUP is currently a hot research topic.

Fine tuning

The theory of **H** makes many assumptions, and many of these many not hold in practice. Those assumption include same genetic parameters in the genotyped sample as in the complete population, and exixtence of compete data to account for selection bias. Several studies found that

better accuracies and lower biases of GEBV can be achieved by fine-tuningα, β,τ and ω in **H** defined as below:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tau(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix}$$

Generally, α≈0.8 and β=1-α. Also, τ<1 and ω<1. Note that large ω causes H to be non-positive definite, causing slower convergence or even divergence when MME are solved by iteration, while smaller ω will generally cause better convergence. Another mechanism to improve the convergence rate is to remove old pedigrees or even old data altogether, especially when the base population is heterogenous, i.e., parents are missing across generations.

Large number of genotypes

When $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$ are computed explicitly, the cost is cubic with the number of genotypes. Subsequently, costs with > 50-100k genotypes are excessive. Research in several labs (Harris and Johnson, VanRaden, Legarra and Ducrocq) focuses on implementation of ssGBLUP where the inverse is not needed. These methods rely on the fact that the product of **Gq** can be obtained with linear time with respect to the number of genotypes as:

Gq=MDM'q=(M(D(M'q)))

when the products are computed sequentially as products of a matrix and a vector.

Approximate accuracies

The additional accuracy of GEBV due to the genomic information can be expressed in terms of record equivalents (RE). A contribution due to the genomic information from animal j to animal i in terms of RE is approximately (Misztal et al., 2012):

$$RE_{ij} \sim (g_{ij} - a_{22,ij})^2 acc^2$$

As the mean difference between the genomic and pedigree relationship is very small, with SD < 0.04 in poultry as found by Wang and Misztal (2011), individual contribution is very small. However, large differences between the relationships due to mistakes in pedigree or genotypes greatly inflate the RE. Also, the contribution is small from animals with small own accuracy. This means that a genomic prediction benefits much more from genotyping animals with high accuracy (males with many progenies) than from animals with own records only.

The formula above has mainly an educational value as contributions from many animals overlap, creating double counting. For formulas that seem to work well but are more cryptic, see Misztal et al. (2102).

Genome wide associations with ssGBLUP

Much of genomic work especially in academia aims at finding genome-wide associations (GWAS) and therefore solutions of SNP effects are desirable. ssGBLUP can be adapted for GWAS (Wang et al., 2012). This involves the following steps:

1. Convert GEBV to SNP effects
2. Estimate individual SNP variances
3. Incorporate variances in G
4. Possibly recompute GEBV and iterate

Experiences indicate that such methodology is similar in accuracy to BayesB, however, at a muchlower costs. It also allows GWAS with effects that are hard to account for with multi-step methods, such as maternal or random regressions, and complicated models including multiple-trait.

The extra benefit of GWAS is availability of SNP weights. Such weights can be incorporated into **G**, possibly increasing the accuracy for traits with major genes. However, in multiple-trait analysis the genomic relationship matrix needs to be identical for all the traits. In such analysis, major markers can possibly be modeled as separate effects, different for each trait.

Conclusions

The single-step methodology provides for easy incorporation of genomic data into a genetic evaluation. The methodology is ready for single-breed evaluations with up to 50k genotypes; extensions to multiple breeds and larger number of genotypes are under way. Also, tools for approximating accuracies and for exploring genome wide associations are available.

The realized accuracy of a genomic evaluation is dependent on many factors, including the quality of genomic data and the structure of the population. In commercial setting, the use of any genomic methodology needs to be accompanied by on-going validation and subsequent troubleshooting if the results fall short of expectations. See FAQ on genomic selection (Misztal, 2011). Such work needs the expertise of well-trained quantitative geneticists who possess substantial quantitative skills while being able to understand the peculiarity of the molecular data and the real architecture of traits.

Acknowledgements

This study was partially funded by the Holstein Association, Smithfield Premium Genetics, Pig Improvement Company, and AFRI grants 2009-65205-05665 and 2010-65205-20366from the USDA NIFA Animal Genome Program.

References

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743:752.

Aguilar, I., I. Misztal, S. Tsuruta, G. R. Wiggans and T. J. Lawlor. 2011a. Multiple trait genomic evaluation of conception rate in Holsteins. J. Dairy Sci. 94:2621-2624.

Aguilar, I., I. Misztal , A. Legarra , and S.Tsuruta. 2011b. Efficient computation of genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128(6):422-428.

Chen, C. Y., I. Misztal, I. Aguilar, S. Tsuruta, T. H. E. Meuwissen, S. E. Aggrey, T. Wing, and W. M. Muir. 2011. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: an example using broiler chickens. J. Animal Sci. 89:23-28.

Chen, C. Y., Misztal, I., Aguilar, I., Legarra, A. & Muir, W. M. (2011). Effect of different genomic relationship matrices on accuracy and scale. SJ. Anim. Sci. 89 : 2673-2679.

Christensen, O. F., P. Madsen P, B. Nielsen, T. Ostersen, and G. Su, G. 2012. Single-step methods for genomic evaluation in pigs. Animal (in print)

Christensen, O., and M. Lund. (2010) Genomic prediction when some animals are not genotyped. Gen. Sel. Evol., 42, 2.

Forni, S., I. Aguilar, I. Misztal (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43: 1.

Goddard, M. E., T. H. E. Meuwissen, and B. J. Hayes. 2010. Genomic selection in farm animal species – lessons learnt and future perspectives. Proc. 9th World Cong. Genet. Appl. Livest. Prod. Paper 0701.

Habier, D., R. L. Fernando and J. C. M. Dekkers. 2007 The impactof genetic relationship information on genome-assisted breedingvalues. Genetics 177: 2389–2397.

Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci., 93, 1243-1252.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain and M. E. Goddard,2009a Invited review: Genomic selection in dairy cattle: progressand challenges. J. Dairy Sci. 92: 433–443.

Hayes, B. J., H. D. Daetwyler, P. J. Bowman, G. Moser, B.Tier et al.,2009b Accuracy of genomic selection: comparing theory andresults. Proc. Assoc. Advmt. Anim. Breed. 17: 352–355.

Hayes, B. J., P. M. Visscher and M. E. Goddard, 2009c Increasedaccuracy of artificial selection by using the realized relationshipmatrix. Genet. Res. 91: 47–60.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656-4663.

Meuwissen, T. H., B. J. Hayes and M. E. Goddard, 2001 Predictionof total genetic value using genome-wide dense marker maps. Genetics157: 1819–1829.

Misztal, I. 2007. Shortage of quantitative geneticists in animal breeding - Editorial. J. Anim. Breed. Genet. 124: 255–256.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. J. Dairy Sci. 92:4648-4655.

Misztal, I. 2011. FAQ for genomic selection - Editorial. J. Anim. Breed. Genet. 128: 245–246.

Misztal, I., S. Tsuruta, I. Aguilar, A. Legarra, and T.J.Lawlor. 2011. Approximation of Genomic Accuracies in Single-Step Genomic Evaluation. Interbull Bul. 43.

Muir, B., Ragavendran, A., Tosa, G. J. M., T. H. E. Meuwissen, I. Misztal, M. Groenen, T. Wing, R. Okimoto, A. Vereijken, H. Cheng. 2012. Genomic Selection in Poultry, Results with Broilers and Comparison with Traditional BLUP. Proc. Plant Anim. Genome Meet. W591.

Simeone, R., I. Misztal, I. Aguilar, and Z. Vitezica. 2012. Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. J, Anim. Breed. Genet.
129(1):3–10.

Tsuruta, S., I. Aguilar, I. Misztal, and T. J. Lawlor. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J. Dairy Sci. 94:4198-4204.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. Bias in Genomic Predictions for Populations Under Selection. Genet. Res. Camb. 93:357–366.

VanRaden, P. M. (2008) Efficient methods to compute genomic predictions. J. Dairy Sci., 91, 4414-4423.

VanRaden, P.M., C. P. Van Tassell, G.R. Wiggans, T. S. Sonstegard,R. D. Schnabel et al., 2009 Invited review: Reliability of genomicpredictions for North American Holstein bulls. J. Dairy Sci. 92: 16–24.

VanRaden, P, M. 2012. Avoiding bias from genomic pre-selection in converting daughter information across countries. Interbull Bull. (in print)

Wang, H., and I. Misztal. 2011. Comparisons of numerator and genomic and relationship matrices. J. Anim. Sci. 89(E-Suppl. 1):163.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94(2):73-83.