SNPMace – A meta-analysis to estimate SNP effects by combining results from multiple countries

M.E. Goddard[1,2], A. Jighly[2], H. Benhajali[3], H. Jorjani[3] and Z. Liu[4]

1. University of Melbourne, Australia

2. Agriculture Victoria, Bundoora, Australia

3. Interbull Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, S-750 07, Uppsala, Sweden

4. IT solutions for animal production (vit), Germany

Abstract

Selection of bulls and cows is increasingly made on genomic estimated breeding values (GEBVs) calculated from their SNP genotypes and the estimated effects of each SNP. To obtain the most accurate GEBVs a large training population of animals with phenotypes and genotypes is needed. For some traits, some breeds and some countries such a large training population is not available. In these cases it would increase the accuracy of GEBVs if information from multiple countries and breeds were combined. This paper describes a meta-analysis to combine SNP effects from multiple countries. A project to test this procedure is under way and, if successful, may result in a new Interbull service.

Interbull context

This paper is supported by the members of the Interbull SNPMace Working Group. The SNPMace WG was established by, and reports to the Interbull Steering Committee. The WG has been established in order to follow and support the development of the methodology in close connection with both research institutions and stakeholders. The SNPMace WG members are: Enrico Santus (Chair), Toine Roozen (Secretary), Mike Goddard, Vincent Ducrocq, Esa Mäntysaari, Zengting Liu, Haifa Benhajali and Hossein Jorjani.

Introduction

A meta-analysis means combining results from multiple experiments or datasets to obtain more accurate estimates of the parameters without combining the raw data. They are widely used to summarize results from medical experiments on the same or similar treatments and recently have been widely used in human genetics. For instance, (Yengo et al. 2018) reported a meta-analysis of genome wide association studies (GWAS) for the trait human height. Most SNPs have a very small association with height and so to estimate this association, with any reliability, requires a very large sample size. By combining results from multiple GWAS, the total sample size is much larger than any individual study making the results much more reliable. Recently a similar meta-analysis of GWAS has been published for stature in cattle (Bouwman et al 2018).

MACE EBVs, calculated by Interbull by combining progeny tests conducted in different countries, are an example of a meta-analysis. They have been of great value to the dairy industry by increasing the accuracy of selection of bulls and allowing comparisons between bulls evaluated in different countries. It would have been possible, in theory, to calculate international EBVs by combining the raw data

from all participating countries and performing one analysis. However, this was not possible or necessary. Individual countries did not wish to share their data with a foreign country and accurate international EBVs could be calculated by the meta-analysis known as MACE (Schaeffer 1994) which uses as input the EBVs calculated within each country rather than the raw data.

However, selection of bulls and cows is increasingly made on the basis of genomic estimated breeding values (GEBVs) calculated from SNP genotypes and the estimated effects of these SNPs. (An equivalent method uses the SNP genotypes to calculate a genomic relationship matrix). Breeding companies and dairy farmers want these GEBVs to be as accurate as possible and this is equivalent to making the estimated SNP effects as accurate as possible. We could combine GEBVs across countries in a similar manner to combining EBVs based a progeny test as is done in the current GMACE service, but this limits the accuracy of these international GEBVs. The reason for this limitation is that the genetic correlation between the same trait (e.g. milk yield) measured in different countries is less than 1. Consequently, a GEBV calculated in country A is less accurate when used in country B than it would be in country A. One way to overcome this limitation is to apply the prediction equation of country B to the genotypes of the bull from country A. However, this approach is limited by the accuracy of the prediction equation for milk yield in country B. This Bulletin suggests an alternative approach in which estimated SNP effects are combined by combining estimated SNP effects from multiple countries so that all prediction equations have higher accuracy than that achieved using data from a single country.

The accuracy of GEBVs depends on the proportion of genetic variance explained by the SNPs and the accuracy of the estimated SNP effects. The accuracy of estimated SNP effects can be increased by increasing the size of the training population, using a single step analysis, sequence data, a non-linear or Bayesian estimation method and possibly in the future using information about the function of polymorphic sites in the genome.

In some countries, some breeds will soon have very large training populations for some traits (e.g. Holsteins in USA for milk yield). However, there will be many breeds, countries and traits where the size of the training population will always be limiting the accuracy. This applies to important traits such as feed efficiency, to countries with small dairy industries and to numerically small breeds. By combining data over countries and breeds we could increase the accuracy of SNP effects and therefore GEBVs.

It is also possible to combine the raw data from multiple countries and perform a single analysis to estimate SNP effects in each country and EBVs for each country. This is done by Interbull in the InterGenomics service. However, not all countries want to share raw data. This paper describes how a meta-analysis can be used to combine the estimated SNP effects from different countries without the need to exchange raw data or genotypes and describes a project to test the feasibility of Interbull offering a new service to combine SNP solutions across countries.

Methods

An international SNP model

Our multi-trait BLUP model assumes that the effects of a SNP in countries i and j (g_i and g_j) are genetically correlated with the same correlation as the genetic correlation between true breeding values in the different countries. Within country i (i = 1,…,c) the SNP effects are estimated as g_i

where g_i is a vector of estimated national SNP effects of country *i*.

For the sake of simplicity, we assume that the input national SNP effect estimates for country *i* are estimated with a SNP BLUP model (Liu et al., 2016) that would be equivalent to:

$$\mathbf{y}_i = \mu_i \mathbf{1} + \mathbf{Z}_i \mathbf{g}_i + \mathbf{e}_i \quad [1]$$

where $\mathbf{y}_i$ is a vector of phenotypes (deregressed proofs) of reference animals corrected for all but additive genetic effects of an original genomic model; μ_i is a general mean of country *i*; $\mathbf{1}$ is a vector of 1s; $\mathbf{Z}_i$ represents the design matrix for genotypes of reference animals. Genotypic values of reference animals take 3 possible values (VanRaden, 2008): $2-2p_j$, $1-2p_j$ and $0-2p_j$ for genotypes AA, AB or BB, respectively, p_j represents allele frequency of SNP marker *j* (*j*=1, …, *m*); $\mathbf{e}_i$ is a vector of residual effects for the reference animals with a (co)variance matrix:

$$\left[\text{var}(\mathbf{e}_i)\right]^{-1} = \mathbf{R}_i^{-1} = diag\{n_{ik}\sigma_{e_i}^{-2}\} \qquad [2]$$

with $\sigma_{e_i}^2$ representing error variance of country *i* and n_{ik} represents the daughter contribution of reference animal *k* in country *i* based on the reliabilities of the bull and its parents.

Under the SNP BLUP model (Liu et al., 2016) SNP effects are distributed as:

$$\text{var}(\mathbf{g}_i) = \mathbf{B}_i\sigma_i^2 \qquad [3]$$

where

$$\mathbf{B}_i = \frac{1}{\sum_j 2p_{ij}(1-p_{ij})}\mathbf{I} = \theta_i\mathbf{I} \qquad [4]$$

(VanRaden, 2008)

σ_i^2 represents variance of direct genomic values (DGV) of country *i*.

Please note that DGV represents the sum of all SNP effects:

$$\text{DGV}_{ik} = \mathbf{z}_{ik}\mathbf{g}_i \qquad [5]$$

where DGV_{ik} is direct genomic value for animal *k*; $\mathbf{z}_{ik}$ is a row in the design matrix $\mathbf{Z}_i$ corresponding to the animal *k*.

Mixed model equations (MME) can be set up equivalently as if the SNP effects of the country were estimated with:

$$\begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{Z}_i \\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i + \sigma_i^{-2}\mathbf{B}_i^{-1} \end{bmatrix}\begin{bmatrix} \hat{\mu}_i \\ \hat{\mathbf{g}}_i \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{y}_i \\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{y}_i \end{bmatrix} \qquad [6]$$

Note that the general mean μ_i is expressed on the DGV, whereas it is usually expressed in national genomic evaluation on genomic breeding values (GEBV) which is the sum of DGV and RPG.

For the SNPMace model [1], SNP effects from different countries have the following (co)variance matrix:

$$\text{var}\begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_c \end{bmatrix} = \begin{bmatrix} \sigma_1^2\mathbf{B}_1 & \sigma_{12}\mathbf{B}_{12} & \cdots & \sigma_{1c}\mathbf{B}_{1c} \\ & \sigma_2^2\mathbf{B}_2 & \cdots & \sigma_{2c}\mathbf{B}_{2c} \\ & & \ddots & \vdots \\ symm. & & & \sigma_c^2\mathbf{B}_c \end{bmatrix} = \mathbf{G} \qquad [7]$$

and its inverse matrix is:

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}^{11} & \mathbf{G}^{12} & \cdots & \mathbf{G}^{1c} \\ & \mathbf{G}^{22} & \cdots & \mathbf{G}^{2c} \\ & & \ddots & \vdots \\ symm. & & & \mathbf{G}^{cc} \end{bmatrix} \qquad [8]$$

where σ_{i,i^+}^2 is DGV covariance between countries *i* and i^+. In order to guarantee sum of the SNP genetic covariances equal the total additive genetic covariance between the two countries, all the involving countries must code the three possible SNP genotypes in the same way, e.g. AA=2, AB=1 and BB=0.

Similar to the definition of matrix $\mathbf{B}_i$ for country *i*, matrix $\mathbf{B}_{i,i^+}$ for the two countries relies on the assumption that the same set of SNP markers are used in the two countries:

$$\mathbf{B}_{i,i^+} = \frac{1}{\sqrt{\sum_j 2p_{ij}(1-p_{ij})}\sqrt{\sum_j 2p_{i^+j}(1-p_{i^+j})}}\mathbf{I} = \sqrt{\theta_i\theta_{i^+}}\mathbf{I} \qquad [9]$$

It can be seen that matrix $\mathbf{B}_{i,i^+}$ between the two countries is an identity matrix multiplied with a scalar as long as the two countries submit SNP effect estimates derived from the same set of SNP markers. Under the assumption of using the same set of SNP markers by all the *c* countries, the (co)variance matrix of the country SNP effects, Equation [8], become

$$\mathbf{G} = \text{var}\begin{bmatrix}\mathbf{g}_1\\ \mathbf{g}_2\\ \vdots\\ \mathbf{g}_c\end{bmatrix} = \begin{bmatrix}\sigma_1^2\theta_1\mathbf{I} & \sigma_{12}\sqrt{\theta_1\theta_2}\mathbf{I} & \cdots & \sigma_{1c}\sqrt{\theta_1\theta_c}\mathbf{I}\\ & \sigma_2^2\theta_2\mathbf{I} & \cdots & \sigma_{2c}\sqrt{\theta_2\theta_c}\mathbf{I}\\ & & \ddots & \vdots\\ symm. & & & \sigma_c\theta_c\mathbf{I}\end{bmatrix}. \qquad [10]$$

Estimation of SNP effects of the SNPMacemodel

The effects of the SNPMace model [1] are estimated using mixed model equations:

$$\begin{bmatrix}\ddots & & \cdots & & \cdots & \\ & \begin{bmatrix}\mathbf{1}'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{Z}_i\\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i\end{bmatrix} + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{ii}\end{bmatrix} & \cdots & \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Psi}_{ii^+}\end{bmatrix} + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{ii^+}\end{bmatrix} & \cdots\\ & \ddots & \vdots & \vdots\\ & & \begin{bmatrix}\mathbf{1}'\mathbf{R}_{i^+}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}_{i^+}^{-1}\mathbf{Z}_{i^+}\\ \mathbf{Z}_{i^+}'\mathbf{R}_{i^+}^{-1}\mathbf{1} & \mathbf{Z}_{i^+}'\mathbf{R}_{i^+}^{-1}\mathbf{Z}_{i^+}\end{bmatrix} + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{i^+i^+}\end{bmatrix} & \vdots\\ & & & \ddots\end{bmatrix}$$

$$\mathrm{X}\begin{bmatrix}\cdots\\ \begin{bmatrix}\hat{\mu}_i\\ \hat{\mathbf{g}}_i\end{bmatrix}\\ \vdots\\ \begin{bmatrix}\hat{\mu}_{i^+}\\ \hat{\mathbf{g}}_{i^+}\end{bmatrix}\\ \cdots\end{bmatrix} = \begin{bmatrix}\cdots\\ \begin{bmatrix}\mathbf{1}'\mathbf{R}_i^{-1}\mathbf{y}_i\\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{y}_i\end{bmatrix}\\ \vdots\\ \begin{bmatrix}\mathbf{1}'\mathbf{R}_{i^+}^{-1}\mathbf{y}_{i^+}\\ \mathbf{Z}_{i^+}'\mathbf{R}_{i^+}^{-1}\mathbf{y}_{i^+}\end{bmatrix}\\ \cdots\end{bmatrix} \qquad [11]$$

The residual (co)variances between countries *i* and i^+, $\boldsymbol{\Psi}_{ii^+}$, depends on the fact if the two countries use bull MACE phenotypes containing common daughter information in their national genomic evaluations. If the MACE EBV of reference bulls are used in national SNP effect estimation in countries i and i^+, the residual covariance can be defined as:

$$\boldsymbol{\Psi}_{ii^+} = (\mathbf{Z}_i'\mathbf{R}_i^{-\frac{1}{2}})(\mathbf{R}_{i^+}^{-\frac{1}{2}}\mathbf{Z}_{i^+}) \qquad [12]$$

If the two countries use only national phenotypes for their SNP effect estimation, then

$$\boldsymbol{\Psi}_{ii^+} = 0 \qquad [13]$$

The residual covariance between the SNP effects of the two countries, $\boldsymbol{\Psi}_{ii^+}$, depends on the number of common reference bulls used in the two national reference populations and EDC of those common reference bulls (Sullivan 2016). However, the best policy would be for countries to use only local data in calculating

their estimated SNP effects. This parallels the current procedure for MACE.

National data for the SNPMace evaluation

Countries need to submit national SNP effect estimates: g_i, and $\mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i$ for a measure of prediction error (co)variances of the SNP effect estimates. All the participating countries must code two SNP alleles A and B in the same way. Marker allele frequencies of a reference SNP allele, like allele A, as well as the variance of direct genomic values must be provided by the countries for the international SNP effect estimation. Because different genomic models may be used in national genomic evaluations, like the genomic BLUP model (GBLUP) or Bayesian genomic models (Meuwissen et al., 2001), we show below how the countries obtain national SNP effects for the SNPMace evaluation from a genomic model other than the SNP BLUP model.

Converting GEBV of the GBLUP model to SNP effects

Countries may use a GBLUP model, either single-step or multi-step ones, for genomic evaluation. GEBV of the GBLUP model can be converted directly to SNP effects following Liu et al. (2016):

$$\mathbf{g}_i = (1-k)\mathbf{B}_i\mathbf{Z}_i'\mathbf{G}_{rel}^{-1}\mathbf{u}_i^* \qquad [14]$$

where k is proportion of residual polygenic variance in total additive genetic variance, $\mathbf{u}_i^*$ is a vector of GEBV of reference animals, and Zi as defined before (design matrix for genotypes of reference animals) . The reliability values can be obtained from:

$$\mathbf{G}_{rel} = (1-k)\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i' + k\mathbf{A}_i \qquad [15]$$

with $\mathbf{A}_i$ representing pedigree relationship matrix of the reference animals.

SNP effects from the Bayesian genomic models

The SNPMace model [1] makes the same assumption on SNP variances as the SNP BLUP model. Additionally, the SNPMace model assumes the SNP markers explain equal genetic covariance among the SNP markers. The assumption of equal SNP genetic variances may be relaxed by allowing heterogeneous SNP genetic variances, like the Bayesian genomic models (Meuwissen et al., 2001). Likewise, we could also relax the assumption on each SNP contributing equally to the total genetic covariance between any country pair.

The Interbull SNPMace project

To test this method of combining SNP solutions from different countries, Interbull is conducting the Interbull SNPMace project, using data from the Brown Swiss breed in 6 countries. As part of the "InterGenomics" service, Interbull already receives individual animal data from Brown Swiss bulls consisting of genotypes and EBVs, and Interbull carries out an analysis of the combined data to estimate SNP solutions.

In the SNPMace project,

1. the Interbull Centre use the meta-analysis described above to estimate SNP solutions and compare these with those obtained through InterGenomics.
2. Jighly and Goddard at Agriculture Victoria are writing software to perform the meta-analysis.

If the project is successful, Interbull will be able to use this software to offer a service to other breeds and countries that are not in a position to supply individual animal data but would like to have SNP solutions estimated from multiple countries or breeds.

This SNPMace project will run from April 2018 to Oct 2019. Then a decision will be made about the usefulness of this method and whether to develop it further as a service by Interbull.

Future Developments

More accurate EBVs can be obtained by using genome sequence genotypes and non-linear or Bayesian methods to estimate SNP effects. This is especially advantageous when combining data from different breeds. The analysis could then identify sequence variants that are particularly useful and these could be imputed

from routine SNP genotypes or directly genotyped by countries participating in the analysis.

Acknowledgements

The authors would like to thank the Brown Swiss community for providing the data, and the Interbull Steering Committee, Interbull Centre, Sweden, and the Department of Economic Development, Jobs, Transports and Resources (DEDJTR), Australia for their support.

References

Bouwman, A. C., Daetwyler, H. D., Chamberlain, A. J., Ponce, C. H., Sargolzaei, M., Schenkel, F. S., et al. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nature genetics, 50(3), 362.

Liu, Z., M.E. Goddard, B.J. Hayes, F. Reinhardt, & R. Reents, 2016. Technical note: Equivalent genomic models with a residual polygenic effect. J. Dairy Sci. 99:2016-2025.

Meuwissen, T.H.E., B.J. Hayes, & M.E. Goddard, 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Schaeffer, L. R. 1994. Multiple-country comparison of dairy sires. J. Dairy Sci. 77:2671.

Sullivan, P.G. 2016. Using a parameter space to improve GMACE evaluation and reliabilities. Interbull Bulletin: 50:1-10.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. bioRxiv, 274654