Multitrait Across Country Genomic Evaluations for EuroGenomics Countries
—
Daughter information approximation method

Hanni Kärkkäinen[1], Vincent Ducrocq[2], Søren Borchersen[3], Gert P. Aamand[4], Esa A. Mäntysaari[1]
[1] Natural Resources Institute Finland
[2] French National Institute for Agricultural Research (INRA)
[3] 3EuroGenomics Cooperation, [4]Nordic Cattle Genetic Evaluation (NAV)

Abstract

EuroGenomics countries share bull genotypes, enabling multitrait across country SNP-BLUP evaluations using pseudo phenotypes from all countries directly. We have previously shown that such model is feasible and benefits the participants. However, in the EuroGenomics countries there is a vast amount of cow data and cow genotypes that the countries would like to be included into the model, but without sharing the actual genotypes. The Daughter information approximation method is our attempt to accomplish this. Our method would require sharing only the national full reference population SNP-solutions, and the prediction error variances of the already shared bulls. The method is pilot tested by partitioning the currently shared EuroGenomics data into subsets mimicking the shared bulls and the unshared parts of the national reference populations.

Key words: genomic selection, MACE, SNP-BLUP, GBLUP, Newton method

Introduction

EuroGenomics (EG) countries[1] constitute a consortium that shares bull genotypes. Thus, unlike in the Interbull SNPMace project (Liu & Goddard, 2018) it is possible to build a genuine multitrait across country SNP-BLUP evaluation using pseudo phenotypes from all countries directly. We have previously demonstrated and validated the performance of EuroGenomics multitrait across country SNP evaluation, and have shown that such model is feasible and benefits the participants (Kärkkäinen et al., 2019).

Shared bulls comprise, however, only a part of the reference populations in the countries. In particular, countries have genotyped a large number of cows, whose information they would like to include into the evaluation—but without sharing the genotypes. We propose a method to use efficiently the shared genotypes and the information from the full national reference population, with minimum workload for the national evaluation centers. Our Daughter information approximation method (DIA) would require (in addition to currently shared data) only SNP solutions based on the full national reference and the corresponding prediction error variances (PEV) for the shared bulls.

In this study we first describe the data and results from EG multitrait SNP model. Then we present results from a pilot test of the DIA method using the current, shared EG data but by randomly dividing the observations into subsets functioning as the shared and unshared parts of national reference.

[1] Germany (DEU), Nordic countries Denmark, Finland and Sweden (DFS), France (FRA), The Netherlands (NLD), Spain (ESP) and Poland (POL)

Materials and Methods

EuroGenomics Data

The phenotypic data used consists of estimated breeding values (EBV) for protein yield, somatic cell score (scs) and female fertility (cc2) from national genetic evaluations of AI-bulls. These are the same EBVs the countries submit to Interbull evaluation. Also reliabilities of the EBV, effective daughter contributions (EDC) and the trait heritabilities were provided by the countries. The EBVs were deregressed using the EDCs, pedigrees and country-wise heritabilities, and the acquired deregressed proofs (DRP) were used as pseudo phenotypes in the model. The genetic correlations between countries were acquired from Interbull (http://www.interbull.org/ib/maceev_archive).

The SNP genotypes shared by EuroGenomics consortium were received from Nordic Cattle Genetic Evaluations (NAV), where they were imputed into standard set used in NAV Holstein evaluations. After imputation the genotypic data consists of 46,342 segregating biallelic markers with 0,1,2 coding.

Total number of animals with observations in the analyses was 35,188. Majority of the bulls (90% of those with protein record) have daughters only in their country of origin, but there still are more than 1000 bulls with daughters in at least 4 countries.

Multitrait across country SNP-BLUP

The multitrait across country SNP-BLUP model equation takes form

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}\mathbf{g} + \mathbf{e},$$

with $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_c)$, where $\mathbf{y}_i \in \mathbb{R}^{n_i}$ is vector of pseudo phenotypes of country i with n_i records; $\boldsymbol{\mu} = (\mathbf{1}^{n_1}\mu_1, \ldots, \mathbf{1}^{n_c}\mu_c)$, where μ_i is the general mean of country i; $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_c)$, where $\mathbf{Z}_i$ is the genotype matrix of country i, (all countries have the same set of m markers with same 0,1,2 coding); $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_c)$, where $\mathbf{g}_i$ is the vector of marker effects in country i; and finally $\mathbf{e} = (\mathbf{e}_1, \ldots, \mathbf{e}_c)$, where $\mathbf{e}_i$ is the vector of random residual effects of country i individuals.

The (co)variance structure of the marker effects consists of

$$\mathrm{Var}(\mathbf{g}) = \begin{bmatrix} \mathbf{I}^m\sigma_{s_1}^2\theta_1 & \cdots & \mathbf{I}^m\sigma_{1c}\sqrt{\theta_1\theta_c} \\ \vdots & \ddots & \vdots \\ symm. & \cdots & \mathbf{I}^m\sigma_{s_c}^2\theta_c \end{bmatrix} = \mathbf{D},$$

where $\theta_i = 1/\sum_{l=1}^{m} 2p_l(1-p_l)$, with p_l denoting the allele frequency of locus l, $\sigma_{s_i}^2$ is the sire variance of country i and σ_{ii^+} the genetic covariance between countries i and i^+; while $\mathbf{I}^m$ is an identity matrix, and

$$\mathrm{Var}(\mathbf{e}) = \begin{bmatrix} \mathbf{R}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ symm. & \cdots & \mathbf{R}_c \end{bmatrix} = \mathbf{R},$$

where $\mathbf{R}_i = \sigma_{e_i}^2 \mathrm{diag}(1/\mathrm{EDC}_{ij})$, with $\sigma_{e_i}^2 = \sigma_{si}^2(4-h_i^2)/h_i^2$, h_i^2 denoting the heritability and EDC_{ij} the effective daughter contribution of animal j at country i, respectively. The residual covariance between countries is zero, as the estimated breeding values of the countries are based on the national records only. The variance components were considered as known parameters, either given by the countries or Interbull, or estimated prior to the across country SNP-BLUP.

In addition to the basic model, we tested models with 10, 20 or 30% of residual polygenic effect. The polygenic effect was introduced into the model as an additional random variable with zero mean and covariance proportional to the pedigree based numerator relationship matrix.

Validation Method

In order to validate the method, the data was split into learning (reference population) and validation sets by bulls' birth date, so that the youngest 10% of the records at each country were assigned to the validation set. Only records with $R_{DRP_v}^2 \geq 0.5$ were used in the validation (except for Poland cc2 $R^2_{DRP_v} \geq 0.3$, due to limited number of records), and animals with EDC ≥ 10 in at least 10 herds in learning. Validation animals having a record in other country's learning set were removed from the latter.

Animal direct genomic values (DGV) were computed from the SNP solutions ($\hat{\mathbf{g}}_i$) and residual polygene estimates ($\hat{u}_{ij}$) as $\hat{a}_{ij} = \mathbf{z}_{ij}\hat{\mathbf{g}}_i + \hat{u}_{ij}$ for animal *j* in country *i*. Validation reliability was defined as $R^2_v = \left(\mathrm{Cor}(\mathrm{DRP}_v, \mathrm{DGV}_v)\right)^2 / R^2_{\mathrm{DRP}_v}$, and bias *b*1 was tested by a weighted linear regression of DRP_v on predicted DGV_v, using EDC_v as weights (Mäntysaari et al. 2010). The model prediction was compared to both current EuroGenomics practice i.e. using MACE proofs (computed for the study using only EG bulls) for the exchange bulls (VanRaden & Sullivan, 2010), and country-wise single trait SNP-BLUP.

Daughter information approximation method

The approach is based on a concept that, instead of sharing raw geno- and phenotype data, we can combine information from national full reference genomic evaluations in the form of summary statistics from the national evaluations. When we know the SNP solutions and the left hand side (LHS) of the mixed model equations of the national evaluation, we can solve the corresponding right hand side (RHS). SNP solutions with international information can then be acquired by using the national LHS and RHS in a SNP MACE framework.

The previous step is the same as in Liu & Goddard (2018). However, as EuroGenomics countries share the bull genotypes, we could construct the left hand side matrix

$$\begin{bmatrix} \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}_i^{-1}\mathbf{Z}_i \\ \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{1} & \mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}_i^{-1} \end{bmatrix}$$

if we would know the $\mathbf{R}_i^{-1}$, *i.e.* the weights for the bulls (here $\mathbf{D}_i = \mathbf{I}^m \sigma^2_{s_i} \theta_i$ represents the national SNP variance matrix). Now, if each country *i* will provide the prediction error variances

$$PEV_{ij} = \mathbf{z}_j \left[\left(\mathbf{Z}_i' \mathbf{R}_i^{-1} \mathbf{Z}_i + \mathbf{D}_i^{-1} \right)^{-1} \right] \mathbf{z'}_j$$

for the exchanged bulls *j*, we can equate the provided PEVs to

$$PEV_{ij} =: \left[\left(\mathfrak{R}_i^{-1} + \mathbf{G}_i^{-1} \right)^{-1} \right]_{j,j}$$

where $\mathbf{G}_i = \boldsymbol{\mathcal{Z}}_i \mathbf{D}_i \boldsymbol{\mathcal{Z}}_i'$ is the genomic relationship matrix of the shared bulls of country *i* with shared genotype matrix $\boldsymbol{\mathcal{Z}}_i$, and $\mathfrak{R}_i^{-1}$ consists of the weights of the bulls that will, in our restricted set of exchanged animals, lead into the same PEV that the country has computed with all animals (including females) in the national genomic evaluation. The estimated weights $\mathfrak{R}_i^{-1}$ are, in a sense, supposed to counteract the different genotype matrix $\boldsymbol{\mathcal{Z}}_i$ in the limited shared data –situation.

We call the estimated equivalents of EDC as daughter information counts (DIC) to separate them from the EDCs submitted by the countries. The new DIC are estimated as follows: First the new weight matrix for the shared bulls of country *i* is written as $\mathfrak{R}_i^{-1} = \sigma_{e_i}^{-2} \mathrm{diag}(\mathrm{DIC}_{ij})$, where $\sigma_{e_i}^{-2}$ is the inverse of the residual variance of country *i*, and DIC_{ij} the initial daughter information count for animal *j* in country *i*. Next, the DICs are estimated iteratively with a Newton method using the PEVs. The general description of the Newton method is

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},$$

where $f(x)$ is the function whose root is to be found. By considering *dic* as the parameter *x* and (PEV_estimated − PEV_from_countries) as the function, the method corresponds with the general parametrization: in each iteration a new estimate dic_{k+1} is computed as

$$dic_{k+1} = dic_k - \mathbf{C}_k^{-1}(\mathrm{PEV}_k - \mathrm{PEV})$$

where dic_k is the current estimate, PEV_k is the prediction error variance computed as $\mathrm{diag}(\mathrm{LHS}^{-1})$ using the current estimate dic_k, and PEV consists of the PEVs the country has computed with the full national reference population.

Matrix $\mathbf{C}_k$, or the value of the partial derivative of $(\text{PEV}_k - \text{PEV})$ with respect to *dic* at the point dic_k, simplifies into a Hadamard (*i.e.* entry-wise) product $\mathbf{C}_k = \text{LHS}_k^{-1} \circ \text{LHS}_k^{-1}$.

After the iteration, the estimated DIC, the shared genotypes, and the SNP solutions countries would provide are incorporated into a series of country-wise SNP-BLUP models, whose RHS-vectors are then used as RHS in the multitrait across country SNP-BLUP.

Pilot study

The Daughter information approximation method was pilot tested with the available EG protein yield data as follows:

1. The data was partitioned into subsets so, that the validation set was as defined previously, but the remaining 90% of the observations were divided randomly into two equal sized groups; one group functioning as "shared bulls" and the other as unshared part of the national reference population.
2. The "shared" SNP solutions and animal PEVs were computed country by country using the "full national reference", that is, the whole reference data set (i.e. the oldest 90% of the bulls).
3. The PEVs were used to estimate the corresponding DIC with the Newton method.
4. The "shared" SNP solutions, "shared" genotypes and the estimated DIC were used in solving the country-wise RHS.
5. The country-wise RHS were incorporated into a multitrait across country SNP-BLUP.
6. Finally, multitrait DGVs were validated using the youngest 10% of the animals.

The procedure, starting from a random division of the data, was performed 4 times, and the validation reliabilities were averaged over the iterations.

Computational methods

The computations were performed with MiX99 release XI/2017 version 17.1107 (MiX99 Development Team, 2017) and snp_blup_rel version 0.51 (Mäntysaari & Stranden, 2016). The programs were modified to suit our purposes. The variance component estimation was performed with the MTG2 (Lee et al. 2016).

Results & Discussion

The multitrait across country SNP-BLUP evaluations produced on average similar validation reliability than the current EuroGenomics practice (Figure 1), but were slightly less biased (higher b_1).

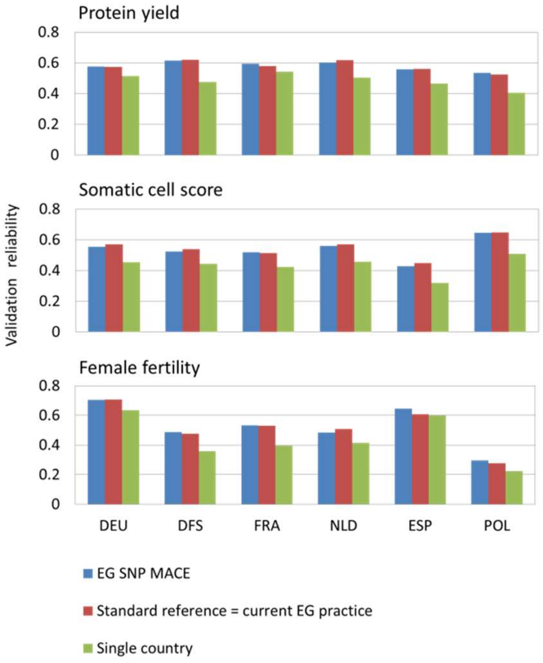

Figure 1. Validation reliability R_v^2 of DGV predicted by EG SNP MACE, by current EuroGenomics practice and by single country SNP-BLUP.

All of the models benefit from residual polygenic component; on average 20% polygenic component seems to be the best choice. Under the EG SNP MACE, the polygenic component increases the average validation reliability from 0.56 to 0.58 for protein yield, 0.53 to 0.54 for somatic cell score and 0.52 to 0.53 for female fertility. The bias diminishes from 0.87 to 0.94, 0.82 to 0.84 and 0.77 to 0.81, respectively.

The estimated variance components did not differ remarkably from the Interbull estimates, and we decided to use the latter ones.

In our pilot study the "full national reference" information in multitrait across country model gave significantly higher accuracy of the DGV (Figure 2) than MT evaluations using "exchange bull" set. Incorporating the additional information indirectly with the Daughter information approximation method improves the accuracy slightly in comparison to the model based solely on the "shared bulls" (red vs. green bar in Figure 2). However, when the extra information is added directly, the improvement is considerably larger (blue bar in Figure 2). The average R_v^2 increased from 0.51 to 0.53 with indirect and to 0.56 with direct addition of extra data.

The problem with direct addition of the full national reference information is that the countries would have to share all of the genotypes and phenotypic data. A possible alleviation to that requirement would be that, in concordance with the Melbourne/Interbull project, the countries would submit the SNP solutions and the left hand sides of the full national genomic evaluation MME. Using these we could solve the right hand sides, and incorporate them to the multitrait across country SNP-BLUP. In theory this should provide the same results than the full model. However, also here the countries would have to submit large amount of data (50,000 × 50,000 matrix, or about 10Gb, for every trait). Also, we do not know how well the method would work in practice with EG data.

On the other hand, in the Daughter information approximation the transfer of information from national evaluations to MT EG evaluations is through SNP solutions. Thus, the difference between "All shared" and DIA (Figure 2.) comes from different weights of information. It can be expected that the difference between methods diminishes when the linkage disequilibrium represented in country-wise LHS of shared bulls becomes proportionally closer to full national reference LHS. Therefore, it is expected that with the larger true full national data DIA will provide a better approximation.

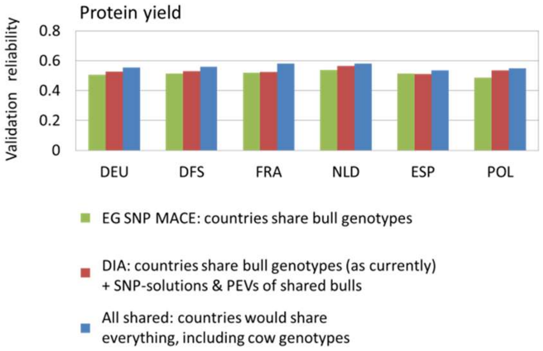

Figure 2. Pilot study validation reliabilities. R_v^2 of DGV predicted by EG SNP MACE mimicking the full national reference populations, by Daughter information approximation and by EG SNP MACE mimicking the shared bulls only.

Conclusions

According to our pilot study, the Daughter information approximation offers a simple way to include the full national reference data into SNP MACE without extensive data exchange. Full reference information increases the accuracy of GEBV over the ones based on "shared" data only. However, including the full national reference information directly would probably improve the evaluation even more, but would require more shared data and more work from the countries.

Acknowledgements

The project is financed by Natural Resources Institute Finland (Luke), French National Institute for Agricultural Research (INRA), EuroGenomics Cooperation and German Livestock Association.

References

Interbull MACE evaluations archive, April 2018. http://www.interbull.org/ib/maceev_archive

Lee, S.H. and van der Werf, J.H.J., 2016. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics 32(9):1420-1422.

Liu, Z. and Goddard, M.E. 2018. A SNP MACE model for international genomic evaluation: technical challenges and possible solutions. *Proc. WCGALP* 11.393.

Kärkkäinen, H., Ducrocq, V., Borchersen, S., Aamand, G.P. and Mäntysaari, E.A. 2019 Multitrait Across Country Genomic Evaluations for Eurogenomics Countries – Research Plan and First Results. Interbull Bulletin (54).

MiX99 Development Team 2017. MiX99: A software package for solving large mixed model equations. Release XI/2017. Natural Resources Institute Finland (Luke). Jokioinen, Finland. URL:http://www.luke.fi/mix99

Mäntysaari, E.A., Liu, Z. and VanRaden, P., 2010. Interbull validation test for genomic evaluations. Interbull bulletin (41), p.17.

Mäntysaari & Stranden, 2016 Mäntysaari, E.A. and Stranden, I., 2016. The software snp_blup_rel: Inside out. October 2016. Interbull document.

VanRaden, P.M. and Sullivan, P.G., 2010. International genomic evaluation methods for dairy cattle. GSE 42:7.