SNPMace - A meta-analysis to estimate SNP effects across countries

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

1. Agriculture Victoria, Bundoora, Australia
2. Interbull Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, S-750 07, Uppsala, Sweden
3. University of Melbourne, Australia
4. IT Solutions for Animal Production (vit), Heinrich-Schroeder-Weg 1, D-27283 Verden, Germany

Abstract

The accuracy of genomic prediction could be improved by combining datasets across countries, but it is not always possible to combine the individual animal data. This project has tested a meta-analysis, called SNPMace, that mimics the combined analysis but requires only summary statistics, such as estimated SNP effects, from participating countries. The method uses the genetic correlation between a trait measured in different countries to produce country specific estimated SNP effects. We tested this method on data from 6 countries on the protein yield of Brown Swiss cattle and on the milk, fat and protein yields of Australian Holstein and Jersey cattle. In both cases the meta-analysis generated estimated breeding values that had a correlation with those obtained by analyzing the raw data in the range 0.99 to 1. The method is implemented in software called MetaGS which also converts data on a subset of SNPs to a common SNP set for analysis.

Key words: SNPMace, international evaluation, genomic, bull, cow, SNP effects, meta-analysis

Introduction

The recent development of high-throughput genotyping techniques makes it possible to genotype large populations with thousands to millions of single nucleotide polymorphisms (SNPs) covering the whole genome in reasonable time and cost (Ding & Jin 2009). Thus, genetic variants affecting different traits can be mapped through their linkage disequilibrium with nearby SNPs. The effect of these loci on traits can be detected using genome-wide association studies (GWAS) which estimate the effect of allele substitution in the population by independently fitting SNPs in a linear regression (between phenotype and genotype) model that takes the population relatedness as a covariate. However, this single SNP regression is not the optimal approach to predict the genetic value of an individual because it ignores the linkage disequilibrium among the SNPs (Wray et al. 2014). In a single SNP regression several SNPs in high LD with each other may all have large effects, but in fact they could all underlie the same causative variant (Udler et al. 2010). Instead, the best way to predict the genetic value or breeding value is to estimate the effects of all SNPs when they are fitted simultaneously. This analysis generates a prediction equation for the performance of new individuals that have been genotyped even if they have no phenotypic records, which was called genomic selection or genomic prediction (Meuwissen et al. 2001). The advantages of genomic prediction have been well empirically demonstrated in plant breeding (Crossa et al. 2014) and animal breeding (Goddard et al. 2011) to select best candidates for different traits, as well as in personalized medicine of complex diseases to predict genetic risk (Abraham & Inouye 2015).

Most SNPs have very small effects on complex traits and so estimating these effects,

with any reliability, requires a very large sample size (Hayes et al. 2009). Often the limited number of individuals, with both phenotypes and genotypes that are available to train the prediction model, limits the accuracy of prediction. By combining data from multiple populations, or studies, the total sample size can be increased making the results much more reliable. Ideally, raw data from different populations can be combined in a single joint analysis to increase the power and accuracy of the association and prediction analyses. However, such analysis is not always feasible because of raw data sharing and privacy restrictions (Tenopir et al. 2011). Alternatively, different dataset holders may agree to share summary statistics from their own data which can allow a meta-analysis to be performed.

A meta-analysis means combining results from multiple experiments or datasets to obtain more accurate estimates of the parameters without combining the raw data (Fleiss 1993). MACE EBVs, calculated by Interbull by combining progeny tests conducted in different countries, are an example of a meta-analysis. Different meta-analysis approaches have been proposed to estimate SNP effects across several datasets and were shown to improve prediction accuracy compared to any of the individual analyses (Bolormaa et al. 2014; Pasaniuc & Price 2017; Maier et al. 2018; Vandenplas et al. 2018). The main drawback to these methods is that most of them depend on approximation, so they suffer from accuracy reduction compared to the joint analysis involving raw datasets. Vandenplas et al. (2018) showed through simulation that their method was as accurate as the joint analysis that combine raw datasets. However, their method is applicable only when the genetic correlation among different experiments is equal to one. The method described by Goddard et al. (2018) overcomes this limitation. A practical problem with carrying out a meta-analysis occurs if there are differences in the SNP lists used in the different experiments.

In this paper, we are testing the multiple best linear unbiased prediction model (multi-BLUP) proposed by Goddard et al. (2018) to improve the accuracy of SNP effect estimates that make use of summary statistics from different datasets instead of utilizing raw data (SNPMace). The method depends on minimal approximation and it almost exactly reproduces a joint analysis that involves all individual raw datasets. Data holders are expected to share the SNP effect estimations, the pairwise linkage disequilibrium between their SNPs, the frequency of the reference SNP allele and the error variance in their populations. We tested our method on two independent datasets, 1) six populations of the Brown Swiss cattle breed with protein yield data and 2) Australian Holstein and Jersey cattle with data for milk, fat and protein yields.

Materials and Methods

Our multi-trait BLUP model assumes that the effects of a SNP in population *i* and *j* ($\boldsymbol{g}_i$ and $\boldsymbol{g}_j$) are genetically correlated with the same correlation as the genetic correlation between true breeding values in the different populations. Within country *i* (*i* = 1,…,c) the SNP effects are estimated as $\boldsymbol{g}_i$ where $\boldsymbol{g}_i$ is a vector of estimated SNP effects of population *i*.

SNP effect estimation in a single population

The input to the meta-analysis are SNP effects estimated within each country. We assume that the input individual SNP effect estimates for country *i* are estimated with a SNP BLUP model (Liu et al., 2016) that would be equivalent to:

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

Where $\boldsymbol{y}_i$ is a vector of phenotypes of the training or reference population corrected for all effects except additive genetic effects explained by the SNPs; $\mu_i$is a general mean of population *i*; $\mathbf{1}$ is a vector of 1s; $\boldsymbol{Z}_i$ represents the design matrix for genotypes of reference individuals. Genotypic values of reference

population take 3 possible values: $2-2p_{ij}$, $1-2p_{ij}$and $0-2p_{ij}$for genotypes AA, AB or BB, respectively (VanRaden, 2008), p_{ij} represents allele frequency of SNP marker j (j=1, …, m) of the population i; $\boldsymbol{e}_i$ is a vector of residual effects for the reference population with a (co)variance matrix:

$$[var(\boldsymbol{e}_i)]^{-1} = \boldsymbol{R}_i^{-1} = diag\{n_{ik}\sigma_{e_i}^{-2}\} \quad [2]$$

with $\sigma_{e_i}^2$ representing the error variance of population i, and n_{ik} representing the effective number of daughters contributing to y_{ik} of reference individual k in population i.

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

$$var(\boldsymbol{g}_i) = \boldsymbol{B}_i\sigma_i^2 \quad [3]$$

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

(VanRaden, 2008)

σ_i^2 represents variance of direct genomic values (DGV) of country i.
DGV represents the sum of all SNP effects:

$$\mathrm{DGV}_{ik} = \boldsymbol{z}_{ik}\boldsymbol{g}_i \quad [5]$$

where DGV_{ik} is breeding value of individual k explained by SNPs; $\boldsymbol{z}_{ik}$ is a row in the design matrix $\boldsymbol{Z}_i$ corresponding to the individual k.
For this model, the mixed model equations for country i are:

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

(Co)variance of SNP effects in different countries

For the multi-BLUP model, SNP effects from different populations have the following (co)variance matrix:

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

where σ_{i,i^+} is DGV covariance between population i and i^+.

Similar to the definition of matrix $\boldsymbol{B}_i$ for population i, matrix B_{i,i^+} for the two populations relies on the assumption that the same set of SNP markers are used in the two populations:

$$\boldsymbol{B}_{i,i^+} = \frac{1}{\sqrt{\sum_j 2p_{ij}(1-p_{ij})}\sqrt{\sum_j 2p_{i^+j}(1-p_{i^+j})}}\boldsymbol{I} = \sqrt{\theta_i\theta_{i^+}}\boldsymbol{I} \quad [8]$$

The (co)variance matrix of the population SNP effects, Equation [7], becomes:

$$\boldsymbol{G} = var\begin{bmatrix}\boldsymbol{g}_1 \\ \boldsymbol{g}_2 \\ \vdots \\ \boldsymbol{g}_c\end{bmatrix} = \begin{bmatrix} \sigma_1^2\theta_1\boldsymbol{I} & \sigma_{12}\sqrt{\theta_1\theta_2}\boldsymbol{I} & \cdots & \sigma_{1c}\sqrt{\theta_1\theta_c}\boldsymbol{I} \\ & \sigma_2^2\theta_2\boldsymbol{I} & \cdots & \sigma_{2c}\sqrt{\theta_2\theta_c}\boldsymbol{I} \\ & & \ddots & \vdots \\ symm. & & & \sigma_c\theta_c\boldsymbol{I} \end{bmatrix}. \quad [9]$$

and its inverse matrix is:

$$\boldsymbol{G}^{-1} = \begin{bmatrix} \boldsymbol{G}^{11} & \boldsymbol{G}^{12} & \cdots & \boldsymbol{G}^{1c} \\ & \boldsymbol{G}^{22} & \cdots & \boldsymbol{G}^{2c} \\ & & \ddots & \vdots \\ symm. & & & \boldsymbol{G}^{cc} \end{bmatrix} \quad [10]$$

Estimation of SNP effects in the SNPMace model

The effects of the SNPMace model are estimated using the following mixed model equations:

$$\begin{bmatrix} \boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i + \boldsymbol{G}^{ii} & \cdots & \boldsymbol{G}^{ii^+} \\ & \ddots & \vdots \\ symm. & & \boldsymbol{Z}_{i^+}'\boldsymbol{R}_{i^+}^{-1}\boldsymbol{Z}_{i^+} + \boldsymbol{G}^{i^+i^+} \end{bmatrix} \times \begin{bmatrix}\hat{\boldsymbol{g}}_i \\ \vdots \\ \hat{\boldsymbol{g}}_{i^+}\end{bmatrix} = \begin{bmatrix}\boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{y}_i \\ \vdots \\ \boldsymbol{Z}_{i^+}'\boldsymbol{R}_{i^+}^{-1}\boldsymbol{y}_{i^+}\end{bmatrix} \quad [11]$$

The terms in equation 11 can all be derived from the individual country analyses. Data holders need to submit the components that allow building equation [11] which are,

1) SNP effect estimates $\boldsymbol{g}_i$;

2) $\boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i$ for a measure of prediction error (co)variances of the SNP effect estimates;

3) Marker allele frequencies of a reference SNP allele like allele A; and

4) the variance of direct genomic values. All the participating data holders must code the two SNP alleles A and B in the same way, so they end with equivalent $\boldsymbol{g}_i$ estimations and $\boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i$ matrices across populations.

It is not necessary for data holders to submit multiple $\boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i$ matrices if they phenotyped different sets of individuals for different traits. This matrix represents the LD structure in the population. Thus, it should be similar for different traits except for a difference in scale. We recommend each data holder to calculate a single $\boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i$ using all genotyped individuals even if they do not have phenotypes. In this case, they are required to submit the number of individuals used to generate $\boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i$ (a) as well as the number of phenotyped individuals for trait i (p_i). The $\boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i$matrix can be rescaled with the number of phenotypes to avoid overestimating the magnitude of populations with missing phenotypes using the following equation:

$$\text{Rescaled } \boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i = p_i/\alpha \times \boldsymbol{Z}_i'\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i \quad [12]$$

Handling different sets of SNP markers between populations

Here we propose a method to account for different SNP datasets used by different populations. In this method we expand the list of SNPs to include all SNPs used by any of the participating countries. Equation 6 shows how the right hand sides (RHS) of the mixed model equations for each country can be obtained from the left hand sides that the country provides i.e the design matrix and the SNP solutions. However, these RHS are missing for SNPs not used by that country, so we impute the missing RHS as follows. We assume that, due to LD among the SNPs, the genotypes for the complete set of SNPs ($\mathbf{Z_c}$) on the bulls used by country i are related to the genotypes for non-missing SNPs ($\mathbf{Z_i}$) by $\mathbf{Z_c = Z_i T}$ where **T** is an i x c matrix where i = number of SNPs used by country i and c = number of all SNPs. **T** can be calculated by:

$$\mathbf{T = (Z_i'Z_i)^{-1} Z_i'Z_c} \quad [13]$$

This requires a set of animals with both the missing and non-missing SNPs recorded. This could be done within country i or by using a reference set of animals at the Interbull Centre. Then the RHS for the missing SNPs can be calculated by

$$\mathbf{Z_c' R^{-1}y = T'Z_i'R^{-1}y} \quad [14]$$

and the LHS by

$$\mathbf{Z_c'R^{-1} Z_c = T'Z_i'R^{-1}Z_iT} \quad [15]$$

This gives all the necessary inputs for equation 11.

After the complete equations have been solved yielding prediction equations for each country based on the complete SNP set, the solutions for the SNP set of country i can be obtained by

$$\mathbf{g_i = Tg_c} \quad [16]$$

where $\mathbf{g_c}$ are the SNP solutions for country i based on the complete SNP set and $\mathbf{g_i}$ are the solutions if only the country i SNP set are to be used.

In order to make $(\mathbf{Z_i'Z_i})^{-1}$ invertible, the number of individuals in the reference population should be larger than the number of SNPs. While it is impossible to attain such large reference population for most organisms on the whole genome scale, we recommend applying it within each LD block separately and setting all off diagonal or inter LD block elements to zero. In the validation analysis of this paper, we calculated multiple submatrices for **T** for every 50 adjacent SNPs on the same chromosome.

Implementation and testing

The SNPMace model and the imputation methods have been implemented in the software MetaGS. The software was multi-threaded and was highly optimized for computational time and memory use. MetaGS was written in C++ and the conjugate gradient method was used to solve the mixed linear model for large datasets. For small datasets,

another solver that calculates the inverse of the left-hand side matrix was implemented in the software for better performance. MetaGS calculates the $\mathbf{Z_i'R^{-1}Z_i}$ matrix and saves it in a smaller size binary format.

Data

i. Australian Holstein and Jersey data genotyped with 50k SNP chip.

After filtering for minor allele frequency, the dataset involved 40850 SNPs. Three traits were considered, milk yield, milk fat and milk protein yields. The reference population included 1071 Jersey and 4105 Holstein bulls born before 2010. The validation set contained 107 Jersey and 522 Holstein bulls born after 2010. The raw data was analysed using MTG2 (Lee & Van der Werf 2016) for comparison with SNPMace. The correlation between Holstein and Jersey bulls was estimated using MTG2 and fitted in the SNPMace model to accurately redundant both analyses. The correlation was 0.54, 0.36 and 0.33 for milk, fat and protein; respectively.

ii. Brown Swiss protein yield data obtained from six countries (Switzerland, Germany, France, Italy, Slovenia and the United States) genotyped also with the 50k SNP chip.

The number of individuals was 1748, 2490, 167, 1275, 227 and 482, all reference bulls respectively from the six countries and the number of SNPs was 45473. Official correlations from the Interbull April 2018 MACE evaluation (https://interbull.org/static/web/proddoc1804r.pdf) of the Brown Swiss data (Table 5) was fitted in the SNPMace model as well as DenseMap software that runs a multi-BLUP analysis on the raw data.

We used the Holstein population to calculate the accuracy of the imputation method (2) as it contains a large number of individuals. We randomly split the population into reference with 3000 individuals with all SNPs (40850) and validation with 1627 individuals and randomly selected 40000 SNPs. The size of the binary $\mathbf{Z_i'R^{-1}Z_i}$ file for each country was equal to ~7.8Gb which can be easily transferred from different data holders to Interbull.

We calculated the $\mathbf{T}$ matrix from the reference (equation 13) and used it to impute the RHS that includes the missing 850 SNPs (equation 14) and their predicted values in the $\mathbf{Z_i'R^{-1}Z_i}$ matrix (equation 15) of the validation set. The accuracy was estimated as Pearson correlation between predicted values and true values.

Results and Discussion

Comparing the SNPMace model to the joint analysis of the raw data using MTG2 on the Australian data resulted in very similar SNP effect estimations. The correlation between both effects ranged between 0.98 to 0.99 for milk yield, fat and protein traits for both Holstein and Jersey breeds (Table 1, 2 and 3). For the Holstein breed, the correlation of both multi-trait models with the single-trait model were also high (between 0.96 to 0.98) but slightly lower than the correlation of multi-trait models. In contrast, the correlation of the multi-trait models with the single-trait model applied to the Jersey breed were lower, which ranged between 0.66 and 0.83 (Table 1, 2 and 3). This is a result of the relatively large Holstein population (4105 bulls) compared to the Jersey reference population (1071 bulls) which reduced the magnitude of the Jersey's input on the Holstein population. By contrast the small jersey population benefits from the information coming from the larger Holstein population.

The prediction accuracies on the validation set were also very comparable between the SNPMace model and the joint raw analysis. They both showed similar or slightly higher prediction accuracy than the single-trait analysis when calculating GEBVs using the same breed SNP effects (Table 4). However, when using breed SNP effects to calculate the GEBVs of the other breed, the prediction accuracies of both multi-trait models were

much higher than the accuracies of the single-trait model with increases ranged from 0.19 to 0.53 (Table 4). The accuracy of predicting all missing SNPs (850 SNPs) in the $\mathbf{Z_i'R^{-1}Z_i}$ matrix using equation 15 was equal to (0.95±0.07), while the accuracy of predicting missing effects in $\mathbf{g_i}$was equal to 0.93.

Testing the SNPMace model on six populations of the Brown Swiss breed also resulted in a very good concordance with the joint multi-BLUP model. The full analysis required around 12 minutes using 20 processors and the iterative solver required 523 iterations to converge. The correlation of SNP effects between both models ranged between 0.95 and 0.96 (Table 5). The correlation for the direct genomic values of the reference individuals between SNPMace and the joint analysis ranged between 0.997 and 0.999, while their correlations with the single trait model ranged between 0.985 and 0.999 (Table 6). The full analysis (including reading inputs and writing outputs) for the six populations required only 12 minutes using MetaGS software of which only 10 minutes were required by the linear solver.

The SNPMace model should be theoretically more robust and accurate compared to any previously published genome-wide meta-analysis. The model is flexible enough to fit complex traits with high genotype × environment interactions. Contrary to other models, that estimates a single effect value per SNP, SNPMace estimates a SNP effect for each country using the genetic correlation among different populations and their linkage disequilibrium structures. Thus, in the worse scenario in which the correlations among all populations are equal to zero, the output effects will be exactly the same as the input effects. Moreover, it provided a comprehensive mathematical method that can accurately synchronises different datasets to maximize the number of analysed SNPs.

For many years Interbull has combined information from different countries through MACE to provide more accurate EBVs for the dairy industry to use in selection of dairy bulls. In the genomics era there is a benefit in combining information across countries to make GEBVs more accurate. This could be achieved by each country providing Interbull with individual animal data including SNP genotypes. The methodology and software described in this paper provides an alternative in which countries do not have to supply individual animal data but only summary statistics such as estimated SNP effects. The results presented here show that this method (SNPMace) yields almost exactly the same GEBVs as if the individual animal data had been combined.

The method, as currently implemented, is best suited to combining data from different populations of the same breed. Although, we demonstrate its use for combining Holstein and Jersey SNP effects, it is not well suited to combining breeds because with only 50,000 SNPs the linkage disequilibrium between causal variants and SNPs is likely to differ between the breeds. In the future it is hoped to extend the method to multiple breeds.

Conclusion

The SNPMace model is able to duplicate the multi-trait BLUP analysis obtained by combining the raw data. The metaGS software implements this analysis and also converts the SNP set of each country to a common set.

Acknowledgements

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

References

Abraham, G., & Inouye, M. (2015). Genomic risk prediction of complex human disease and its clinical application. Current opinion in genetics & development, 33, 10-16

Bolormaa, S., Pryce, J. E., Reverter, A., Zhang, Y., Barendse, W., Kemper, K., et al. (2014). A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLoS genetics, 10(3), e1004198

Crossa, J., Perez, P., Hickey, J., Burgueno, J., Ornella, L., Cerón-Rojas, J., et al. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity, 112(1), 48-60

Ding, C., & Jin, S. (2009). High-throughput methods for SNP genotyping. In Single Nucleotide Polymorphisms (pp. 245-254). Humana Press, Totowa, NJ

Fleiss, J. L. (1993). Review papers: The statistical basis of meta-analysis. Statistical methods in medical research, 2(2), 121-145

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of animal breeding and genetics, 128(6), 409-421

Goddard, M. E., Jighly, A., Benhajali, H., Jorjani, H., Liu, Z. (2018). SNPMace – A meta-analysis to estimate SNP effects by combining results from multiple countries. Interbull Bulletin: 54:1-6.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science, 92(2), 433-443

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

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

Maier, R. M., Zhu, Z., Lee, S. H., Trzaskowski, M., Ruderfer, D. M., Stahl, E. A., et al. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nature communications, 9(1), 989

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

Pasaniuc, B., & Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics, 18(2), 117-127

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., et al. (2011). Data sharing by scientists: practices and perceptions. PloS one, 6(6), e21101

Udler, M. S., Tyrer, J., & Easton, D. F. (2010). Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. Genetic epidemiology, 34(5), 463-468

Vandenplas, J., Calus, M. P., & Gorjanc, G. (2018). Genomic prediction using individual-level data and summary statistics from multiple populations. Genetics, 210(1), 53-69

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

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., & Middeldorp, C. M. (2014). Research review: polygenic methods and their application to psychiatric traits. Journal of Child Psychology and Psychiatry, 55(10), 1068-1087

Tables

Table 1. Correlations of SNP solutions for milk yield between single trait analysis (ST), multi-trait analysis (MT) and SNPMace model. Jersey SNP solutions above diagonal, Holstein below diagonal and standard deviation of SNP solution (Holstein/Jersey) on the diagonal.

Milk yield	ST	MT	SNPMace
ST	(0.386 /0.198)	0.66	0.69
MT	0.96	(0.394 /0.284)	0.98
SNPMace	0.97	0.99	(0.399 /0.285)

Table 2. Correlations of SNP solutions for milk fat between single trait analysis (ST), multi-trait analysis (MT) and SNPMace model. Jersey SNP solutions above diagonal, Holstein below diagonal and standard deviation of SNP solution (Holstein/Jersey) on the diagonal.

Milk fat	ST	MT	SNPMace
ST	(0.012 /0.007)	0.79	0.77
MT	0.97	(0.012 /0.009)	0.98
SNPMace	0.98	0.99	(0.012 /0.009)

Table 3. Correlations of SNP solutions for milk protein between single trait analysis (ST), multi-trait analysis (MT) and SNPMace model. Jersey SNP solutions above diagonal, Holstein below diagonal and standard deviation of SNP solution (Holstein/Jersey) on the diagonal.

Milk prot	ST	MT	SNPMace
ST	(0.0094 /0.0057)	0.81	0.83
MT	0.98	(0.0094 /0.0066)	0.99
SNPMace	0.98	0.99	(0.0096 /0.0066)

Table 4. Prediction accuracy for milk yield, fat and protein yields using single trait analysis (ST), multi-trait analysis (MT) and SNPMace model.

		ST		MT		SNPMace	
		Jer	Hol	Jer	Hol	Jer	Hol
Yield	Jer	0.52	0.32	0.5	0.46	0.53	0.5
	Hol	0.05	0.51	0.49	0.52	0.46	0.52
Fat	Jer	0.34	0.18	0.37	0.36	0.37	0.36
	Hol	0	0.52	0.31	0.53	0.3	0.53
Protein	Jer	0.55	0.15	0.54	0.39	0.54	0.4
	Hol	0.08	0.48	0.39	0.53	0.4	0.53

Table 5. Above diagonal, correlations between populations that was used in the SNPMace and the multi-BLUP models; diagonal and below diagonal, the correlation of SNP effect estimations using SNPMace and multi-BLUP models.

Pop	NoBulls	CHE	DEA	FRA	ITA	SVN	USA
CHE	1748	**0.95**	0.81	0.89	0.81	0.82	0.92
DEA	2490	0.95	**0.96**	0.82	0.81	0.86	0.85
FRA	167	0.94	0.94	**0.96**	0.81	0.83	0.86
ITA	1275	0.93	0.94	0.94	**0.96**	0.82	0.81
SVN	227	0.94	0.95	0.95	0.95	**0.96**	0.83
USA	482	0.93	0.94	0.95	0.94	0.95	**0.95**

Table 6. Correlation of DGVs calculated using ST, MT and SNPMace analyses for protein yield in BSW.

Country	ST/MT	ST/SNPMace	MT/SNPMace
CHE	0.995	0.997	0.998
DEA	0.996	0.997	0.998
FRA	0.985	0.99	0.997
ITA	0.998	0.999	0.999
SVN	0.998	0.999	0.999
USA	0.988	0.992	0.996