Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.53

Genomic prediction across populations, using pre-selected markers and differential weight models

Biaty Raymond[1,2§], Aniek C. Bouwman[1], Chris Schrooten[3], Jeanine Houwing-Duistermaat[4] and Roel F. Veerkamp[1]

[1]Animal Breeding and Genomics, Wageningen University and Research, P.O. Box 338, 6700 AH Wageningen, The Netherlands

[2]Biometris, Wageningen University and Research, 6700AA Wageningen, The Netherlands

[3]CRV BV, P.O. Box 454, 6800 AL Arnhem, The Netherlands

[4]Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Leiden, The Netherlands

[§]Corresponding author

Email addresses:

BR: biaty.raymond@wur.nl

ACB: aniek.bouwman@wur.nl

CS: chris.schrooten@crv4all.com

JHD: j.duistermaat@leeds.ac.uk

RFV: roel.veerkamp@wur.nl

Introduction

The utilization of genomic prediction (GP) in numerically small populations is limited due to the lack of a large reference sets (Goddard 2009; VanRaden, 2008). Potentially, information can be utilised across populations. In practice however, it has been shown that the use of information across populations results in little to no benefit (Calus et al., 2014; Erbe et al., 2012; Hayes et al., 2009; Kachman et al., 2013) or even negative prediction accuracies (Calus et al., 2014; Hayes et al., 2009). Some studies suggest that with increasing marker density, linkage disequilibrium (LD) between quantitative trait nucleotides (QTNs) and markers will increase, and GP across populations will consequently improve (De Roos et al., 2008; Kizilkaya et al., 2010). Others suggested that focussing on the QTNs themselves might improve GP (Boichard et al., 2012; Hoze et al., 2014).

Our hypothesis is that GP across populations can be improved when: 1) the effects of known QTNs are separated from those of other markers and are properly weighted 2) a multi-trait modelling approach is used, in which the genetic correlation between populations serves as a measure of how much one population's information can contribute to the accuracy of selection in another population. Therefore, the objective of this study was to test alternative models for across and multi-breed GP that appropriately utilise prior information on marker causality.

Material and methods

Phenotype and marker data.

Estimated breeding values (EBVs) for stature and number of daughters were available for 735 Jersey bulls from New Zealand and 5,553 Dutch Holstein bulls. Data was provided by CRV BV (Cooperative Cattle Improvement Organization, Arnhem, the Netherlands). Stature EBVs for all bulls were deregressed to obtain deregressed proofs (DRPs) according to the method proposed by Calus et al., (2016). The output from this procedure were the DRPs and effective daughter contributions (EDCs). Animals with EDCs of zero were removed, resulting in 595 New Zealand Jersey bulls and 5,503 Dutch Holstein bulls in the final dataset. The EDCs were used as weights for the DRPs in subsequent analyses.

Two sets of marker data were used. The first set was labelled as top markers' and comprised of 357 markers that were considered to be causal based on the result of a large multi-breed meta-GWAS analysis using imputed whole genome sequence data (Bouwman et al., 2015). From the list of markers with effect significantly different from zero $P < 5e\text{-}8$ in the meta-GWAS study, only the markers that uniquely contributed to the proportion of explained variance were selected (Yang et al., 2012). The second set of markers was those present on the Bovine 50k chip and not present in the top markers. This second set, hereby designated as 50k, comprised of 48,912 markers. For both marker sets, markers were kept when the minor allele frequency is equal to or greater than , where N is the number of genotyped individuals. This threshold was set within each breed.

Statistical models

Across breed genomic relationship matrices (GRMs) were calculated for all the animals according to the first method of VanRaden (2008). Relationships were computed in the calc_grm software (Calus, 2013). Three GRMs were formed using only the 50k markers (GRM50k), only the top markers (GRMtop) and the 50k and top markers combined into a single GRM (GRMall). Four different GREML models were fitted in ASReml (Gilmour et al., 2009) as follows:

Model 1 used a single trait animal model with the 3 GRMs fitted one at a time.

(1)

Where is a vector containing stature DRPs, is the mean of the DRPs, is a vector of additive genetic effects, is the design matrix that links to the DRPs in and is a vector containing the residuals. Both and are assumed to be normally distributed as and , where is the GRM, the genetic variance, the residual variance and is a diagonal matrix that contains the EDCs, used as weights for the DRPs in the model.

Model 2 used a single trait animal model with GRM50k and GRMtop fitted simultaneously

(2)

All model components are as described under model 1.

Model 3 used a multi-trait animal model with the 3 GRMs fitted one at a time, considering stature in Holstein (H) as a different trait as stature in Jerseys (J).

(3)

where and

Model 4 used a multi-trait animal model with GRM50k and GRMtop fitted simultaneously. The model was:

(4)

where ,

Models 1 and 2 were implemented in a i) within breed prediction scenario (WBP), using only Jersey bulls as reference animals. All the jersey bulls were randomly split into 5 sets and these were used one time each as validation set in a 5-fold cross validation scheme. ii) an across breed prediction scenario (ABP) using Holstein bulls as reference population and Jersey bulls as validation population. The multi-trait models 3 and 4 were implemented in. iii) multi-breed prediction scenario (MBP), using a reference population made up of both Holstein and Jersey bulls and a validation population of only Jerseys. A 5-fold cross-validation scheme was also used to obtained the genomic estimated breeding values of Jerseys as described in WBP. Estimation of the proportion of genetic variance captured within a breed was done using all bulls available within the breed, we used data on all bulls for that specific breed. All animals from the two breeds were used for estimating . In all models, accuracy of prediction was computed as the correlation between the genomic estimated breeding values (GEBVs) of jersey bulls and their DRPs.

Results

Variance components

In general, more genetic variance was explained in the Holstein breed compared to the Jersey (Table 1). When fitted alone, the top markers accounted for 75% of the total genetic variance in Holstein and 49 to 51% in Jersey. However, when fitted simultaneously with the 50k, the top markers captured only 26% and 22% of the total genetic variance in Holstein and Jersey respectively.

In the multi-trait analysis between Holstein and Jersey breeds using the 3 different GRMs, the proportion of genetic variance explained in Holstein and jersey breeds were not higher than those explained in the single trait models. the lowest estimate for (0.25) was obtained when only the 50k was fitted (Table 1). The estimate was almost double (0.45) when only the top markers were fitted. The estimate of for the top markers increased to 0.89 when 50k and top GRMs were fitted together.

Accuracy of prediction

Using solely Holstein bulls in the reference population to predict Jersey bulls GEBV resulted in low accuracies ranging from 0.06 to 0.25 (Figure 1). Although using top markers alone was more accurate than fitting only 50k (0.21 vs 0.06), the accuracy of prediction was further improved when the model fitted top markers and 50k simultaneously (0.25). Unlike in the across breed scenario, fitting the top markers alone in the within and multi-breed prediction scenarios resulted in lower accuracies than the 50k. In both within and multi-breed GP however, model 4 consistently gave the highest accuracy (0.42 and 0.45 respectively), as was in the across breed scenario.

Discussion

Across breed GP can be seen as an approach that circumvents the challenge of small reference population in numerically small breeds. Across breed GP does not take into account the differences in population structure between breeds (De Roos et al., 2009; De Roos et al., 2008; Zhong et al., 2009), most importantly, the differences in LD structure and minor allele frequency. Because of differences in LD structure for example, the estimate of markers effects in the reference and the validation breeds will be different (Hill & Robertson, 1968), thus resulting in low and accuracy of across breed prediction. This explains why in our study,

using only the 50k for ABP resulted in low and a very low accuracy of prediction (0.06). The challenge that remains therefore is to develop an approach that is able to minimize the effect of differences in LD structure on the accuracy of across breed GP.

Our approach for reducing the effect of LD structure differences was to focus directly on the potential causal (top) markers for the trait. Our expectation was that the effects of these markers will be more consistent across breed and less affected by the differences in LD structure, thus result in higher and accuracy. Our results agreed with this expectation (Figure 1). The main limitation of this approach was that the top markers alone did not explain the total genetic variance for stature within the breeds (Table 1). Because of this limitation, we tested other models in which we fit both the top markers and the 50k simultaneously, however giving them different weights (model 2 for ABP and model 4 for MBP). The idea is that these models are able to focus uniquely on the top markers, devoid of influence from the 50k. Moreover, the 50k markers should pick up the remaining genetic variance not explained by the top markers, such that they do not end up in the residual. In all the prediction scenarios, this approach resulted in the highest accuracy of prediction (Figure 1). The difference between model 2 and model 4 is that in model 2, is assumed to be 1 for both the top markers and the 50k, while in model 4, are estimated for both the top markers and the 50k. Furthermore, the estimated implicitly weights the contribution of Holstein information to Jerseys, hence the higher accuracy in model 4 compared to 2.

Although the estimated for the top markers in model 4 increased to 0.89, the corresponding increase in the accuracy of prediction is only marginal compared to fitting only the top markers or top plus 50k markers (Figure 1). This is because the proportion of total genetic variance captured reduced to 26% and 22% in Holsteins and Jerseys respectively. In effect, the covariance between breeds which is a product of and variances captured remains about the same. Moreover, the within breed accuracy for the Jerseys was already relatively high. Our results indicate that for across population GP, it is beneficial to use a model that is able to isolate the effects of potential causal markers and differentially weight them. Furthermore, it helps to weight information across populations by the in a multi-trait model approach, rather than naively assuming a of 1 between populations.

Conclusion

Based on the results of this study, it is feasible to use information from numerically larger breeds like Holstein to improve the accuracy of selection in numerically small breeds under two main conditions. i) animals from the small breed must be included themselves in the reference population together with those from other breeds. ii) the prediction model should take into account the differences in causality between the markers used by for example, separating causal and non-causal markers in separate GRMs as was done in this study.

Acknowledgements

The authors want to acknowledge CRV and the 1000 bull genomes consortium for providing the data. The authors want to also thank the Netherlands Organisation of Scientific Research (NWO) and the Breed4Food consortium partners Cobb Europe, CRV, Hendrix Genetics, and Topigs Norsvin for their financial support.

References

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M.-N., Boscher, M. Y., . . . Journaux, L. (2012). Genomic selection in French dairy cattle. *Animal Production Science, 52*(3), 115-120.

Bouwman, A. C., Pausch, H., Govignon-Gion, A., Hoze, C., Sanchez, M., Boussaha, M., . . . Guldbrandtsen, B. (2015). *Meta-analysis of GWAS of bovine stature with> 50,000 animals imputed to whole-genome*

sequence. Paper presented at the Annual Meeting of the European Association for Animal Production.
Calus, M. (2013). calc_grm–a programme to compute pedigree, genomic, and combined relationship matrices. *Animal Breeding and Genomics Centre, Wageningen UR Livestock Research*.
Calus, M., Vandenplas, J., Ten Napel, J., & Veerkamp, R. (2016). Validation of simultaneous deregression of cow and bull breeding values and derivation of appropriate weights. *Journal of Dairy Science*.
Calus, M. P., Huang, H., Vereijken, A., Visscher, J., ten Napel, J., & Windig, J. J. (2014). Genomic prediction based on data from three layer lines: a comparison between linear methods. *Genetics Selection Evolution, 46*(1), 57. doi: 10.1186/s12711-014-0057-5
De Roos, A., Hayes, B., & Goddard, M. (2009). Reliability of genomic predictions across multiple populations. *Genetics, 183*(4), 1545-1553.
De Roos, A., Hayes, B. J., Spelman, R., & Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. *Genetics, 179*(3), 1503-1512.
Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., . . . Goddard, M. E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J Dairy Sci, 95*. doi: 10.3168/jds.2011-5019
Gilmour, A. R., Gogel, B. J., Cullis, B. R., & Thompson, R. (2009). *ASReml User Guide Release 3.0*. Hemel Hempstead: VSN International Ltd.
Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica, 136*. doi: 10.1007/s10709-008-9308-0
Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., & Goddard, M. E. (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution, 41*(1), 1.
Hill, W., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics, 38*(6), 226-231.
Hoze, C., Fritz, S., Phocas, F., Boichard, D., Ducrocq, V., & Croiseau, P. (2014). *Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds.* Paper presented at the 10ème World Congress of Genetics Applied to Livestock Production.
Kachman, S. D., Spangler, M. L., Bennett, G. L., Hanford, K. J., Kuehn, L. A., Snelling, W. M., . . . Pollak, E. J. (2013). Comparison of molecular breeding values based on within- and across-breed training in beef cattle. *Genetics Selection Evolution, 45*(1), 30. doi: 10.1186/1297-9686-45-30
Kizilkaya, K., Fernando, R. L., & Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotype. *J Anim Sci, 88*. doi: 10.2527/jas.2009-2064
VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J Dairy Sci, 91*. doi: 10.3168/jds.2007-0980
Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., . . . Loos, R. J. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics, 44*(4), 369-375.
Zhong, S., Dekkers, J. C., Fernando, R. L., & Jannink, J.-L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics, 182*(1), 355-364.

Table 1: Proportion of total genetic variance explained by each genomic relationship matrix within Dutch Holstein and New Zealand Jersey bulls and the genetic correlation (between the two breeds (standard errors in parentheses). The within breed estimates were obtained using a univariate model and a single breed's data, while the multi-breed estimates were obtained using a bivariate model in which the data of both breeds were fitted simultaneously.

Model with:	Holstein*	Jersey**
	Within breed estimates	
GRM50k	0.97 (0.00)	0.76 (0.05)

GRMtop	0.75 (0.02)	0.49 (0.06)	
GRMall	0.97 (0.00)	0.76 (0.05)	
GRMs 50k & top	0.70 (0.02) & 0.26 (0.02)	0.57 (0.07) & 0.22 (0.07)	
	Multi-breed estimates		
GRM50k	0.96 (0.00)	0.76 (0.05)	0.25 (0.17)
GRMtop	0.75 (0.02)	0.51 (0.06)	0.45 (0.11)
GRMall	0.97 (0.00)	0.76 (0.05)	0.33 (0.17)
GRMs 50k & top	0.70 (0.02) & 0.26 (0.02)	0.57 (0.07) & 0.22 (0.07)	0.28 (0.20) & 0.89 (0.13)

* Values estimated using all Holstein bulls in the study and Jersey data were masked
**Values estimated using all Jersey bulls in the study and Holstein data were masked

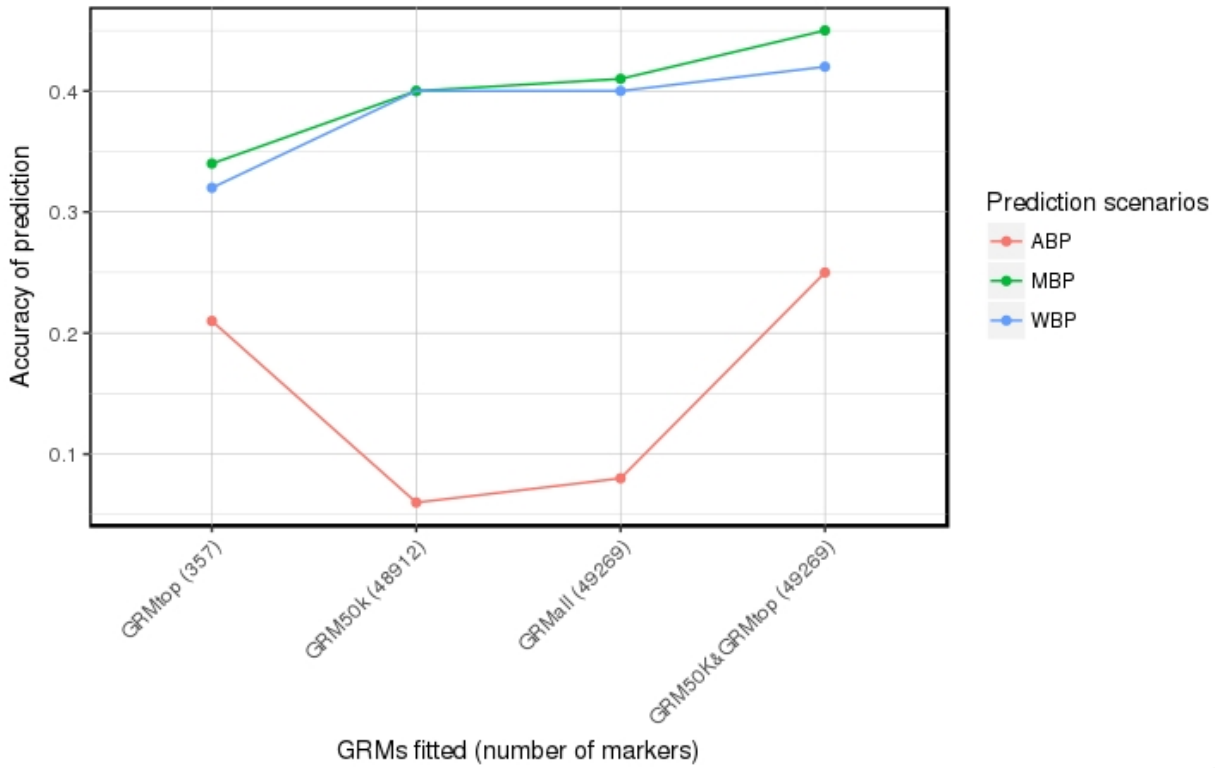

Figure 1: Accuracy of genomic estimated breeding values (GENVs) for Jersey bulls in a within breed (WBP), across breed (ABP) and multi-breed (MBP) scenario using the different genomic relationship matrices (GRMs). Accuracy was computed as the correlation between GEBVs of jerseys and their deregressed proof for stature. GRMtop was calculated based on the genotypes of the top markers, GRM50K based on the genotypes of 50k markers and GRMall based on the genotypes of the 50k and top markers combined.