Contributed paper

USING SELECTED SEQUENCE VARIANTS TO IMPROVE GENOMIC PREDICTION OF HEAT TOLERANCE IN DAIRY CATTLE

E.K. Cheruiyot[1,2], M. Haile-Mariam[1], B.G. Cocks[1,2], I.M. MacLeod[1] and J.E. Pryce[1,2]

[1]Agriculture Victoria, Department of Jobs, Precincts and Regions, Bundoora, VIC 3083, Australia
[2]School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia

SUMMARY

Genomic breeding values for heat tolerance were first developed and released to the Australian dairy industry in 2017, to allow farmers to select animals that better tolerate hot and humid conditions. It is desirable to improve the reliability of these genomic predictions to help accelerate the genetic improvement for this trait. Whole-genome sequence data may contain causative mutations, or variants in high linkage disequilibrium with causal mutations for traits. This study investigated the potential improvements in the accuracy of genomic prediction for heat tolerance when adding informative markers to the 50k industry SNP panel used routinely by DataGene for Australian dairy genomic evaluations. We selected informative sequence variants from a genome-wide association study (GWAS) of heat tolerance phenotypes of 20,623 Holstein cows (each cow with ~15 million imputed sequence variants) and augmented the 50k SNP panel with these SNPs for genomic prediction using a Holstein bull reference (N = 3,323) and Holstein cow validation set (N = 8,484). The accuracy of genomic prediction of heat tolerance for reduction in milk, fat, and protein yield under hot and humid conditions increased by 0.1%, 4%, and 6% units, respectively when informative markers were integrated with 50k SNP data. Since genetic gain is linearly related to EBV accuracy, this lift in accuracy is important for driving the genetic improvement of heat tolerance.

INTRODUCTION

Heat tolerance is the ability of an animal to maintain production and reproductive performance under hot and humid conditions. The desire to breed for heat tolerance is growing worldwide due to the increasing effect of global warming on animal production. Considerable research has been conducted so far in many countries, including Australia, where the first breeding values for heat tolerance were released to the dairy industry in 2017 (Nguyen *et al*. 2017).

Since genetic gain is linearly related to the accuracy of estimated breeding values (EBVs), even a small lift in the accuracy of the heat tolerance EBV is important to the dairy industry. Besides increasing the size of the reference population, one way to boost the accuracy is to increase the density of markers used for genomic predictions. However, increasing the marker set from lower density SNP panels to whole-genome sequence have, in most cases, yielded limited, or no appreciable increase in the accuracies for various traits in cattle (e.g., Van Binsbergen *et al*. 2015). A promising alternative, in which a boost of accuracy has been realized in previous reports (e.g., Moghaddar *et al*. 2019), has been to augment standard industry SNP panels (i.e., 50k or 600K arrays) with a small set of informative or causal mutations for a trait. To fully maximize predictions, this approach requires careful selection of informative markers. Thanks to the 1000 Bull Genomes project (Hayes and Daetwyler 2019), it is now possible to use this sequence database to impute genotyped animals up to whole genome sequence. This may facilitate accurate selection of highly informative variants for use in genomic predictions, especially for complex traits such as heat tolerance.

In this study, we selected informative variants for heat tolerance from a genome-wide association study (GWAS) using milk production records of 20,623 Holstein cows, each having over 15 million

imputed sequence variants. We then investigated the accuracy of prediction when sets of these selected variants were added to the standard industry 50k SNP array, by training the prediction in a bull reference set, and validating it in an independent set of Holstein cows.

MATERIALS AND METHODS

Phenotypes. The phenotypes used in this study were obtained from DataGene (DataGene Ltd., Melbourne, Australia; https://datagene.com.au/) and included test-day milk, fat, and protein yields for Holstein cows and bulls, collected from dairy herds between 2003 and 2017 that were matched with climate data (daily temperature and humidity) obtained from weather stations across Australia's dairying regions. The distribution of dairy herds and weather stations, data filtering, and the calculation of environmental covariate (i.e., temperature-humidity index or **THI**) used in this work were described in our earlier studies (Nguyen *et al*. 2016, Cheruiyot *et al*. 2020).

Calculation of heat tolerance phenotypes for cows and bulls. The rate of decline (slope) in milk, fat, and protein yield due to heat stress events was estimated using reaction norm models as described by Cheruiyot *et al*. 2020. In these models, data on milk, fat, or protein yield were adjusted for fixed effects, including herd test day, year season of calving, parity, age at calving, jointly for parity and DIM, and jointly for stage of lactation and THI. Random effects fitted in the model included a random regression on a linear orthogonal polynomial of THI, where the intercept represents the level of mean milk yield and the linear component represents the change in milk yield (slope) due to heat stress for each cow (i.e., trait deviations (**TD**)) and a residual term. Slope solutions for each bull's daughters were averaged to obtain slope traits for bulls (i.e., daughter trait deviations (**DTD**)).

Genotypes and study design. Two genotype data sets were available: 50k SNP array and ~15 million imputed whole-genome sequence variants. The number of Holstein animals with genotypes and heat tolerance phenotypes were 29,107 ♀/3,323 ♂. We split the Holstein cows into two: 1) QTL discovery set (N = 20,623; comprising older cows born before 2013) for selecting informative markers for heat tolerance, and 2) genomic prediction validation set (N = 1,223; young cows born after 2012). We used Holstein bulls as a training set for genomic prediction. We ensured that none of the cows in the QTL discovery set were daughters of the bulls in the training set to avoid parent-daughter pairs between the two datasets to minimise close genetic relationships.

QTL discovery analysis and selection of informative SNPs. We performed single-trait GWAS analysis to test associations between individual SNP and cows' slope traits (milk, fat, and protein) using GCTA software (Yang *et al*. 2011). The models used for analyses are described by Cheruiyot *et al*. (https://www.biorxiv.org/content/10.1101/2021.02.04.429719v1.full).

Following the GWAS, we selected informative variants defined as 'top SNPs' for each slope trait as follows: for SNPs passing the GWAS threshold of $-log_{10}(p\ value) = 2$; we chose the most significant SNP from within each 100 kb window and sliding 50 kb to the next window along each chromosome. We then removed one SNP of any pair of the selected 'top SNPs' in strong LD ($r^2 > 0.95$).

Genomic prediction. We used BayesR (Erbe *et al*. 2012) to estimate prediction accuracies for 50k SNP panel and compared the resulting accuracies with those obtained from adding 'top SNPs' to the 50k SNP set (i.e., 50k + 'top SNPs') using BayesRC method (MacLeod *et al*. 2016). The BayesR model fitted to the training bulls (N = 3,323) for 42,572 variants from 50k SNP panel was: $y = X\beta + Zg + Wv + e$, where **y** = vector of heat tolerance slope phenotypes; **X** = design matrix; **β** = vector of fixed effect solutions; **Z** = design matrix relating phenotypes to GBV; g = vector of GBV ~ $N(0, \mathbf{I}\sigma_g^2)$, where σ_g^2 is the additive genetic variance for the trait; **W** = design matrix of SNP genotypes; **v** = vector of SNP effects, modelled to have four possible normal distributions corresponding to zero, small, medium and large effects, respectively; **e** = vector of residual errors

$N(0, \mathbf{E}\sigma_e^2)$, where **E** is a diagonal matrix calculated as $diag(1/w_i)$, with w_i being a weighting factor for *i*th sire calculated based on the available number records following Garrick *et al*. (2009).

We then used the BayesRC method to analyse 50k + 'top SNPs' dataset; an extension of the BayesR model that allows pre-allocation of variants to 2 or more classes (MacLeod et al., 2016) and hence a different posterior mixture distribution within each class if the class is enriched for informative SNPs. In our case the SNPs from 50k array (42,572) were allocated to class I and the selected 'top SNPs' to a separate class II, because the latter may be enriched with causal mutations for heat tolerance. For both BayesR and BayesRC models, we performed five MCMC replicate chains, each with 40,000 iterations of which 20,000 were discarded as burn-in for all the traits. We ran the analysis for 2 random validation sets of 600, and 623 Holstein cows.

Calculating accuracy of genomic prediction. For each of the three validation cow sets (described above), the accuracy of prediction was calculated as: $Accuracy(Val_i) = \frac{r_{GBV,phen}}{\sqrt{h^2}}$, where Val_i = Holstein cow validation set; $r_{GBV}, phen$ = correlation of GBV and phenotypes (i.e., slope traits); h^2 = genomic heritability calculated for each trait using variance component estimates of Holstein cows (N = 29,107) for 50k SNP array (45,504 SNPs) data based on –reml option of GCTA software (Yang *et al*. 2011). The bias of prediction was assessed as the regression coefficient of the phenotypes (pre-corrected for fixed effects) on the GBV for animals in the validation set.

RESULTS AND DISCUSSION

In this study, we used a large dataset of Holstein cows (N = 20,623) to select informative markers from a GWAS and then tested them for increased genomic prediction of heat tolerance phenotypes.

The genomic heritability estimates (± standard errors) for the heat tolerance milk (**HTMYslope**), fat (**HTFYslope**) and protein (**HTPYslope**) yield slope traits from Holstein cows that used to calculate the accuracy of predictions were 0.23 ± 0.01, 0.21 ± 0.01, and 0.20 ± 0.01, respectively. The number of informative markers for heat tolerance (i.e., 'top SNPs') selected from GWAS ($p < 0.01$) was highest for HTPYslope (9,633) followed by HTFYslope (9,352), and HTMYslope (9,207) traits. Similarly, the total number of markers used in the BayesRC analyses (i.e., 50k + top SNPs) were 51,750, 51,894, 52,168, for HTMYslope, HTFYslope and HTPYslope traits, respectively. We chose a cut-off of $p < 0.01$, which is comparatively relaxed, to capture both markers with small and large effect sizes for heat tolerance.

Figure 1 shows the accuracy and bias of genomic predictions in the Holstein validation cows. For the BayesR model using only 50k SNP data, we found the highest accuracy of prediction for HTFYslope (0.49 ± 0.01), followed by HTMYslope (0.49 ± 0.01) and HTPYslope (0.39 ± 0.01). The bias across all study traits was > 1.0 (Figure 1) indicating 'deflation' or under prediction, meaning less variance among predicted than observed values.

When the selected 'top SNPs' were added to the standard 50k SNP array and analysed using the BayesRC model, we found a consistent increase in the prediction accuracy across all the traits with values of 0.001, 0.04, and 0.06 for HTMYslope, HTFYslope and HTPYslope traits, respectively (Figure 1). This increase in accuracy is notable for HTFYslope and HTFYslope traits and likely to be associated with the pre-selected markers (potentially functionally linked with heat tolerance) and the method used (BayesRC). The bias of prediction for BayesRC was comparable that for BayesR. In this study, we investigated the potential benefits of sequence variants selected from a single breed (Holsteins) on the accuracy of genomic predictions for the same breed (within-breed prediction). The value of sequence variants selected in across-breed population (combined Holsteins and Jersey) on genomic prediction of other breeds (Jersey and crossbred cattle) will be investigated in a further study.

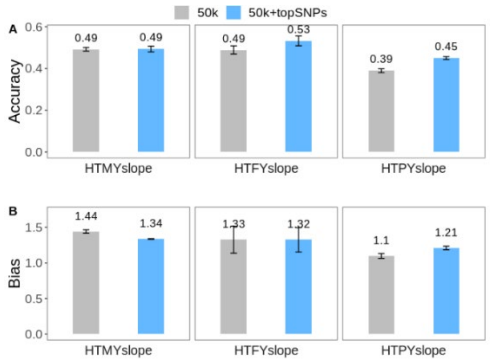

Figure 1. Accuracy of genomic prediction in Holsteins cows based on BayesR (50k; grey) and BayesRC (50k+topSNPs; blue) models for heat tolerance milk (HTMYslope), fat (HTFYslope), and protein (HTPYslope) yield slope traits. Vertical lines are the standard errors of prediction estimated from 2 random validation sets of 600, and 623 Holstein cows

CONCLUSION

Overall, our results show that the accuracy of genomic prediction for reduction in milk, fat, and protein yields under hot and humid conditions can be improved by 0.1%, 4%, and 6% units, respectively when selected informative sequence variants are added to the industry-implemented 50k SNP panel.

ACKNOWLEDGEMENTS

The study was supported by DairyBio (Melbourne, Australia), funded by Dairy Australia, The Gardiner Foundation and Agriculture Victoria. We are grateful to the 1000 Bull Genomes project for access to the sequence data and thank Dr Bolormaa Sunduimijid (Agriculture Victoria Research) for imputation of sequence genotypes. Thanks to DataGene Ltd and Australian dairy farmers for the phenotype and genotype data.

REFERENCES

Cheruiyot, E. K., T. T. T. Nguyen, M. Haile-Mariam, B. G. Cocks, M. Abdelsayed, and J. E. Pryce. (2020) *J Dairy Sci* **103**:2460.

Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, C. Reich, B. Mason, and M. Goddard. (2012) *J Dairy Sci* **95**:4114.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. (2009) *Genet Sel Evol* **41**:55.

Gilmour, A., B. Gogel, B. Cullis, S. Welham, and R. Thompson. (2015) ASReml user guide release 4.1 structural specification. Hemel hempstead: VSN International Ltd.

Hayes, B. J. and H. D. Daetwyler. (2019) *Annu Rev Anim Biosci* **7**:89.

MacLeod, I., P. Bowman, C. Vander Jagt, M. Haile-Mariam, K. Kemper, A. Chamberlain, C. Schrooten, B. Hayes, and M. Goddard. (2016) *BMC genomics* **17**:144.

Moghaddar, N., M. Khansefid, J. H. van der Werf, S. Bolormaa, N. Duijvesteijn, S. A. Clark, A. A. Swan, H. D. Daetwyler, and I. M. MacLeod. (2019) *Genet Sel Evol* **51**:72.

Nguyen, T. T., P. J. Bowman, M. Haile-Mariam, J. E. Pryce, and B. J. Hayes. (2016) *J Dairy Sci* **99**:2849.

Nguyen, T. T., P. J. Bowman, M. Haile-Mariam, G. J. Nieuwhof, B. J. Hayes, and J. E. Pryce. (2017) *J Dairy Sci* **100**:7362.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. (2011) *Am J Hum Genet* **88**:76.

Van Binsbergen, R., M. P. Calus, M. C. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. (2015) *Genet Sel Evol* **47**:1.