GENOMIC PREDICTION AND CANDIDATE GENE DISCOVERY FOR DAIRY CATTLE TEMPERAMENT USING SEQUENCE DATA AND FUNCTIONAL BIOLOGY

I.M. MacLeod[1], P.J. Bowman[2], A.J. Chamberlain[1], C.J. Vander Jagt[1], H.D. Daetwyler[1,2], B.J. Hayes[3] and M.E. Goddard[1,4]

[1] Agriculture Victoria, Centre for AgriBioscience, Bundoora, VIC, 3083 Australia
[2] School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3083 Australia
[3]Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia, QLD, 4067 Australia
[4]Faculty of Veterinary and Agricultural Sciences, University of Melbourne, VIC, 3010 Australia

SUMMARY

Dairy cow temperament is a complex trait affecting both animal and human welfare. Using Bayesian methods, differential gene expression and sequence variant annotation, we increased the accuracy of genomic prediction for temperament compared to using only HD genotypes. Candidate genes for temperament overlapped with genes associated with human neuropsychiatric disorders. More generally, the results indicate that for complex traits, we could make further gains in the accuracy of genomic prediction from access to more specific knowledge of functional biology. This study demonstrates a practical approach to use imputed sequence genotypes and functional biology to improve the accuracy of genomic prediction.

INTRODUCTION

Since the time of cattle domestication some 10,000 years ago there has been continuous genetic selection for animals of docile temperament (excepting animals bred for combat). In dairy cattle, good temperament is critical for animal welfare as well as human safety because of the daily interaction between cattle and agricultural technicians carrying out tasks such as milking and semen collection. Dairy cattle temperament is a polygenic trait with low to moderate heritability (Visscher and Goddard 1995). Given the intensive selection pressure for docility, we hypothesise that a significant proportion of the segregating variants that affect temperament will be relatively rare and recent. If this is the case, it is likely that for candidate gene discovery and genomic prediction there would be an advantage in using sequence variants rather than high density (HD) SNP chips. The reason for this is that SNP on commercial arrays are chosen to be common variants and are therefore not in strong LD with rare variants which are much more common in sequence data.

This study had three aims: 1) to use sequence variants to improve the accuracy of genomic prediction for temperament, 2) to use differential gene expression and functional annotation as a biological prior to increase the accuracy of genomic prediction, 3) to discover candidate genes affecting dairy cattle temperament.

MATERIALS AND METHODS

Phenotypes & Genotypes. Australian dairy cow milking temperament is routinely scored by farmers on a scale of 1 to 5 (where 1 is good and 5 is bad) and phenotypes are processed by DataGene for use in national dairy cattle evaluation. For this study, DataGene provided temperament phenotypes pre-corrected for herd-year-season for Holstein (7,354), Jersey (3,224) and Australian Red (103) animals, including records on 7,343 cows, and 3,338 bulls with progeny test of $\geq$ 20 daughters. Phenotypes were expressed as trait deviations for cows and daughter trait deviations for bulls (mean=-0.20, SD=0.61, min=-2.25, max=3.48) as used for the national dairy cattle evaluations. DataGene

also provided pedigree information. All animals had either real or imputed Illumina 800K BovineHD beadChip genotypes (HD). Subsequently, their genotypes were imputed to sequence variants in all gene coding regions (exons) as well as 5000 bp flanking all known genes. The combined HD and sequence data, "SEQ", was then pruned for SNP pairs in perfect linkage disequilibrium (LD, $r^2 > 0.99$) and for variants with minor allele frequency (MAF) < 0.002 (details in MacLeod *et al.* 2016). After filtering, 994,019 variants remained and the animal genotypes were then centred and scaled to a unit variance. The Australian Reds (all bulls) were used only for validation of genomic predictions. The reference set included all Holstein and Jersey animals.

Statistical models. The data was analysed using the BayesR and BayesRC methods described by Erbe *et al.* (2012) and MacLeod *et al.* (2016) respectively. Briefly the model fitted was:

Temperament = mean + breed-sex group + SNP effects + pedigree + error,

where pedigree was fitted to account for any polygenic genetic variance not explained by the combined SNP effects. To account for heterogeneous error variance associated with cow and bull phenotypes, the residuals were weighted following Garrick *et al.* (2009) and this was implemented in the Bayesian models as described in Kemper *et al.* (2015). Our Bayesian models fit SNP effects jointly as a mixture of four normal distributions with a mean of zero and variance: $\sigma^2_1=0\sigma^2_g$, $\sigma^2_2=0.0001\sigma^2_g$, $\sigma^2_3=0.001\sigma^2_g$ and $\sigma^2_4=0.01\sigma^2_g$, where σ^2_g is the additive genetic variance. All analyses were replicated with 5 MCMC chains, each with 40,000 iterations (20,000 burn-in). The accuracy of genomic prediction was estimated as the correlation between the genomic predictions and phenotypes, and bias was assessed as the regression coefficient of phenotypes on predictions.

The BayesRC approach is very similar to BayesR but incorporates prior biological knowledge in the model. For example, if one or more groups of variants are thought to be more enriched for QTL or causal variants, these can be allocated to a separate variant category *a priori*. In BayesRC, each category is then independently modelled as a mixture of the four BayesR distributions described above, but each starting with equal priors. If a category of variants is found to be enriched for causal variants in the data, this can improve the fit of the model.

Therefore, *a priori* we used independent differential gene expression data measured in 18 bovine tissues (Chamberlain *et al*. 2016), to identify 500 genes that were most highly differentially over-expressed in each of: caudal brain tissue, cerebral brain tissue and adrenal tissue. There was a strong overlap between the top 500 over-expressed genes in each of these three tissues, resulting in a unique set of 1006 genes that we refer to collectively as the "DE" gene set. To further inform the selection of variants for potentially enriched categories, we annotated all non-synonymous coding variants (NSC) associated with the DE genes as well as variants < 50 Kb up- and down-stream of DE genes (REG). We tested four BayesRC models, the first being "DE7" with 7 variant categories (of which 6 used functional annotation):

1) NSC in DE genes overlapping in both caudal and cerebral tissue (N=1617)
2) NSC in DE genes in either caudal or cerebral tissue (N=1447)
3) NSC in the remaining DE genes in adrenal tissue (N=1430)
4) REG flanking DE genes overlapping in both caudal and cerebral tissue (N= 30549)
5) REG flanking DE genes in either caudal or cerebral tissue (N= 28893)
6) REG flanking the remaining DE genes in adrenal tissue (N= 22151)
7) All remaining variants (N= 907932)

"DE2" was the second BayesRC model, where variants in categories 1 to 6 above were combined into one category, and remaining variants to a second category. The third and fourth models, "Random7" and "Random2", had variant categories that matched DE7 and DE2, except that the DE gene set was replaced with a random set of 1006 genes chosen from 24,580 known bovine genes. The BayesR model was run with SEQ or HD genotypes.

RESULTS AND DISCUSSION

The estimated heritability of temperament in the BayesR SEQ model was 0.1 which, although low, indicates that there is still important genetic variation for this trait. Previously, Visscher and Goddard (1995) estimated the heritability of Australian dairy cattle temperament to be 0,2 using only bull progeny test data and a sire model. More recent literature, in Holsteins, report similar heritability estimates to ours for farmer scored temperament (e.g. Stephansen *et al*. 2018). The accuracy of genomic prediction in the Australian Red validation set improved when sequence variants and HD SNP were combined (SEQ) in the BayesR model compared to HD only (Table 1). This may be a result of the sequence variants being in stronger linkage disequilibrium (LD) with causal variants and/or causal variants being included. If it is due to stronger LD, this could reflect the possibility that variants affecting temperament are rare because there has been strong selection pressure for docile temperament in dairy cattle since domestication. Previous studies in cattle for other traits have also shown small improvements from using selected subsets of sequence data compared to 50K or HD SNP genotypes (eg. Brøndum *et al.* 2015; MacLeod *et al.* 2016). However, use of full genome sequence has not yet shown consistent improvement compared to SNP chip genotypes (eg. Calus *et al.* 2016; van den Berg *et al.* 2017). We had therefore pre-selected a subset of sequence variants from gene coding regions and regions adjacent to genes, hoping to capture important missense or regulatory mutations for candidate genes.

In our study, the BayesRC DE7 and DE2 models showed a further small increase in the accuracy of prediction (Table 1). These two models used the same variants as BayesR SEQ, but used prior biology to identify categories of variants that were in or close to genes highly over-expressed in brain or adrenal tissue compared to 17 other tissues (DE genes). Additionally, the DE7 model incorporated a biological prior on variant annotation: non-synonymous coding variants and those that might be regulatory. In the BayesRC Random2 and Random7 models, we replaced the DE gene set with a random set of genes and used this as the prior to group variants into 2 or 7 categories. The accuracy of prediction in the Random2 and Random7 models was lower than the DE2 and DE7 models (Table 1). This lends support to our assumption that genes which are highly expressed in brain and/or adrenal tissue are more enriched for variants controlling dairy cow temperament. However, the level of enrichment for the different variant categories was not very high compared to the random models, suggesting that more specific prior biology is required to better inform the BayesRC model. The accuracy for the Random7 model was slightly lower than the HD. Although this is likely not significant, it could reflect the inclusion of some poorly imputed sequence variants that add noise to the prediction. This could be further tested by constructing random models multiple times. The bias of the predictions suggests a tendency to under-predict genomic breeding value but it is similar across the models.

Our Bayesian methods have previously been demonstrated to be a useful approach for fine mapping genes and mutations that affect complex traits (eg. MacLeod *et al.* 2016). Following our previous study, we used the Bayesian "posterior probability of a variant having a non-zero effect" to detect QTL regions and identify candidate genes. In the BayesR SEQ model, if there is very strong LD across a QTL region, the model will have difficulty distinguishing which variant to prioritise, so the posterior probability will be relatively low and spread across all variants in strong LD. Therefore, to locate candidate gene regions, we summed the posterior probability in windows of 20 SNP, sliding 10 SNP to the next window. We identified 11 known genes in or closest to the top 13 QTL regions genome-wide: NCOA7, GAD2, PDGFD, TMPRSS5, DRD2, IQSEC1, MAOB, PTPRF, SLC25A16, TMCO5A, SNRPB2. The first seven were highly differentially expressed in bovine brain and/or adrenal tissue (Chamberlain *et al*. 2016) in line with our assumption that the DE genes were more likely to be associated with cow temperament than other genes. Furthermore, 10 genes of these 11 overlap candidate

genes or gene families associated with a range of human neuropsychiatric or neurodevelopmental disorders including: schizophrenia, autism, intellectual disability, post-traumatic stress and anxiety (eg. http://atgu.mgh.harvard.edu/~spurcell/genebook/genebook.cgi?user=guest&cmd=overview).

Table 1. Accuracy and bias of genomic prediction in 103 Australian Red bulls using a range of BayesR and BayesRC analytical models

Model[1]	Accuracy	Bias	Increase in accuracy vs. HD
BayesR HD	0.236	1.4	-
BayesR SEQ	0.269	1.6	3.4%
BayesRC DE7	0.289	1.6	5.3%
BayesRC DE2	0.282	1.7	4.6%
BayesRC Random7	0.221	1.3	-1.4%
BayesRC Random2	0.254	1.5	1.8%

[1] See Materials & Methods for acronyms

CONCLUSIONS

This study demonstrates a practical approach to exploiting sequence data and functional biology to improve the accuracy of genomic prediction and for causal gene discovery. It is likely that more specific functional biology would be beneficial for this approach.

ACKNOWLEDGEMENTS

Dairy farmers and DataGene are acknowledged for collection, processing & provision of phenotypes and pedigree data. We are grateful to Dr P Stothard (University of Alberta) for annotation of sequence variants and to the 1000 Bull Genomes Project for access to sequences.

REFERENCES

Brondum R., Guldbrandtsen B., Sahana G., Lund M. and Su G. (2014) *BMC Genomics* **15**: 728.
Calus M.P.L., Bouwman A.C., Schrooten C. and Veerkamp R.F. (2016) *Genet. Sel. Evol.* **48:** 49.
Chamberlain A.J., Vander Jagt C.J., Goddard M.E. and Hayes B.J. (2014) *Proceedings of the World Congress of Genetics Applied to Livestock Production* **Paper 180**.
Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A. and Goddard M.E. (2012) *J. Dairy Sci.* **95:** 4114.
Garrick D.J., Taylor J.F. and Fernando R.L. (2009) *Genet. Sel. Evol.* **41**: 44.
Kemper K.E., Reich C.M., Bowman P., vander Jagt C.J., Chamberlain A.J., Mason B.A., Hayes B.J. and Goddard M.E. (2015) *Genet. Sel. Evol.* **47**: 29.
MacLeod I.M., Bowman P.J., Vander Jagt C.J., Haile-Mariam M., Kemper K.E., Chamberlain A.J., Schrooten C., Hayes B.J. and Goddard M.E. (2016) *BMC Genomics* **17**: 144.
Stephansen R.S., Fogh A. and Norberg E. (2018) *J. Dairy Sci.***101**: 11033.
van den Berg I., Bowman P.J., MacLeod I.M., Hayes B.J., Wang T., Bolormaa S. and Goddard M.E. (2017) *Genet. Sel. Evol.* **49**: 70.
Visscher P. and Goddard M. (1995) *J. Dairy Sci.* **78**: 205.