PITFALLS OF PRE-SELECTING SUBSETS OF SEQUENCE VARIANTS FOR GENOMIC PREDICTION

I.M. MacLeod[1], S. Bolormaa[1], C. Schrooten[4], M.E. Goddard[1,2] and H. Daetwyler[2,3]

[1] AgriBio, Department of Economic Development, Jobs, Transport and Resources, Victoria.
[2] Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Victoria, Australia
[3] School of Applied Systems Biology, La Trobe University, Victoria, Australia
[4] CRV, Netherlands

SUMMARY

Genomic prediction (GP) in farm livestock generally exploits SNP array genotypes. Now it is possible to impute from SNP chip genotypes to whole genome sequence. However, in an industry setting it is impractical to implement GP using millions of sequence variants. Livestock industries are therefore keen to leverage sequence data by selecting subsets of variants to develop custom SNP arrays. In this study we demonstrate that there are potential pitfalls in this approach that can lead to considerable bias in GP and can underestimate the potential advantages of sequence.

INTRODUCTION

Genomic prediction is becoming a popular tool for livestock breeding, and in commercial settings generally exploits SNP array genotypes. Recently, large numbers of animals have been sequenced, enabling imputation to whole-genome sequence for any animal with SNP array genotypes. In theory all imputed sequence variants (> 20 million) could be used for genomic prediction and this should include the causal mutations. However, in practice this is computationally impractical for livestock industries. Furthermore, prediction models that include many millions of imputed sequence variants have not yet increased genomic prediction accuracy relative to SNP array genotypes (van Binsbergen *et al.* 2015; Calus *et al.* 2016). This may be a result of: 1) exacerbated "large p small n" problem leading to an over-saturated model, 2) difficulty in precisely estimating SNP effects due to long distance linkage disequilibrium (LD) and 3) imputation errors. A practical solution is to discover important sequence variants associated with key traits and then design custom SNP arrays that combine the selected variants with SNP from existing commercial arrays (e.g. Wiggans *et al.* 2016). This reduces industry problems associated with large genotype data sets, reduces the "large p small n" analytical issue and increases genotyping accuracy of important sequence variants.

In dairy cattle, several studies have attempted to gain advantage from imputed whole-genome sequence by running a single SNP regression analysis (GWAS) to identify a subset of the most significant sequence variants, and then combining these with lower density SNP array genotypes for genomic prediction (Brøndum *et al.* 2015; van den Berg *et al.* 2016; Veerkamp *et al.* 2016). Similarly, Wiggans *et al.* (2016) demonstrated a small advantage in genomic prediction accuracy by pre-selecting the most informative SNP from high density SNP array genotypes and then using this SNP subset to train the prediction equations. In all these studies, the analysis to select the top variants and their subsequent analysis to train the genomic prediction equations was carried out with the same reference population.

Here, we demonstrate that when pre-selected variants are discovered in the same reference population that is used to train subsequent genomic predictions, this approach can result in significant bias in the predictions. Furthermore, our results suggest that this approach may underestimate potential gains from using subsets of sequence variants in both accuracy and persistency of genomic prediction. We demonstrate that these pitfalls can be avoided by pre-selecting sequence SNP from a population that is independent from the reference population used to train the genomic prediction equations.

MATERIALS AND METHODS

We chose a data set of 21,879 dairy cattle with real genotypes and simulated phenotypes from the same data described in MacLeod *et al.*(2016). Briefly, the genotypes included 2.785 million imputed sequence variants and Illumina 800K Bovine HD beadChip genotypes. Sequence variants included only those in gene coding regions or in putative regulatory regions 5 Kb up- and downstream of genes. After pruning out one of all SNP pairs in perfect LD and SNP with minor allele frequency < 0.002, a total of 994,019 variants remained ("SEQ"). Three trait phenotypes were simulated for all animals by selecting 4000 of these variants to be causal mutations (QTN) with three different genetic architectures and a heritability of 0.6 (details in MacLeod *et al.* 2016). For each trait 3485, 500 and 15 additive QTN effects were sampled from three different normal distributions with a mean of zero and variances of $0.0001\sigma_g^2$, $0.001\ \sigma_g^2$ and $0.01\ \sigma_g^2$ respectively, where σ_g^2 is the additive genetic variance. Breeding values (BV) for all animals were calculated as: $BV_j = \sum_{i=1}^{4000} x_{ij}\alpha_i$, where α_i is the i^{th} QTL effect and x_{ij} represents the i^{th} genotype (coded 0, 1 or 2 for genotypes aa, Aa and AA) of animal *j*.

The animals included 16,133 Holstein, 4861 Jersey and 885 Australian Red breed. The 885 Australian Red and the youngest 584 Holstein were used as two separate validation populations (one distantly related and one closely related). The remaining animals were divided into two separate mixed breed reference sets: Ref1 with 7991 Holstein and 2323 Jersey, and Ref2 with 7558 Holstein and 2538 Jersey. Pedigree records were available for both Ref1 and Ref2. We applied two methods of genomic prediction: GBLUP and BayesR, with the standard model described in MacLeod *et al.* (2016). In the BayesR analyses, variant effects were sampled from four normal distributions with mean of zero and variances as described above for simulated QTN effects. BayesR is a useful method for QTN discovery (e.g. MacLeod *et al.* 2016) so we used BayesR rather than GWAS to identify a subset of putative QTN.

First we undertook QTN discovery separately in Ref1 and Ref2 using the SEQ genotypes (included the surrogate QTN) and then chose the top 500 putative QTN from each analysis. Then we created two custom SNP chips: the first combined the top putative QTN from Ref1with the 50K Illumina BovineSNP50 chip genotypes (Chip_Ref1) and the second combined the top 500 putative QTN from Ref2 with the 50K set (Chip_Ref2). These custom chips were then used for genomic prediction in reference population Ref1. Thus genomic prediction with Chip_Ref1 mimics the approach taken by several recent studies mentioned above: i.e. the QTN discovery population (Ref1) was not independent of the reference population used to train the genomic predictions. In contrast, for Chip_Ref2 the selected putative QTN were discovered in a population (Ref2) that was independent of the one used to train the genomic prediction equations (Ref1). Finally, the two validation populations were used to test accuracy and bias of prediction equations derived from Ref1 with the custom SNP chips as well as the full SEQ, 800K and 50K genotypes. BayesR results are presented as the average of five MCMC chains and results for both GBLUP and BayesR were averaged across the three trait phenotypes (trends being similar). The accuracy of genomic prediction was calculated as the correlation between predicted and true breeding values, and bias was assessed by the regression of the true breeding value on the predicted value.

RESULTS AND DISCUSSION

The accuracy of genomic prediction was highest for SEQ genotypes (Fig 1) as expected because SEQ included all surrogate QTN variants. The relative advantage of SEQ was greater for the Australian Red validation compared to the Holsteins. This reflects the extra precision of the prediction which is only apparent when validation animals are not strongly related to the reference set (Aust. Red breed animals were not in Ref1 or Ref2). The BayesR accuracy was always higher than GBLUP. This was not surprising because we simulated a mixture model with many small effects and a few large effects and Bayesian models are generally superior to GBLUP for this scenario. For BayesR and GBLUP there was an increase in accuracy using either Chip_Ref1 or Chip_Ref2 compared to 50K only. However, for BayesR this advantage was greater for Chip_Ref2 where the putative QTN were discovered in a population that was independent of the reference population that was subsequently used to train the genomic prediction equation. For Chip_Ref2 the accuracy of prediction exceeded the accuracy of the HD 800K and the relative increase was higher for Australian Reds than Holsteins. For GBLUP there was little difference in the accuracy from the two custom chips. However, when we created custom chips by combining the top 5000 SNP from the SEQ analyses with the 50K set, the accuracy of GBLUP markedly improved with Chip_Ref2 compared to Chip_Ref1 (results not shown).

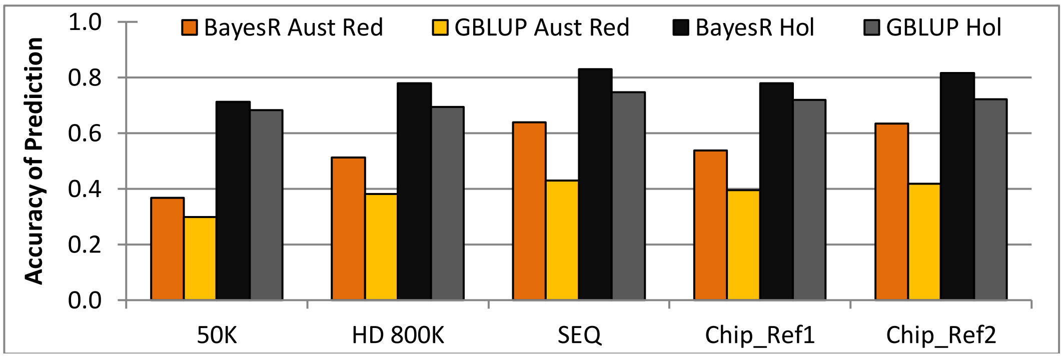

Figure 1. Accuracy of genomic prediction equations trained in Ref1 using a range of SNP genotypes and validated in Holsteins and Australian Reds. SEQ represents ~1million sequence variants, including the SNP chosen as surrogate QTN. Chip_Ref1 is a custom chip of 50K + 500 top putative QTN discovered in the same Ref1, while Chip_Ref2 is a custom chip with 50K + 500 top putative QTN discovered independently in Ref2.

Overall, the bias of genomic prediction (Fig 2) was largest for Chip_Ref1 where the top SNP were discovered in the same set as subsequently used to train the genomic predictions (Ref1). The regression was < 1 which indicates that genomic breeding values were over-predicted. This over-prediction can cause problems for the industry because genomic breeding values would be biased upwards compared to traditional breeding values. We were able to correct the bias (BayesR and GBLUP) in both validation sets, by using custom Chip_Ref2. We also investigated the proportion of variance explained by SNP in each analysis, and found that this variance was considerably over-estimated in the case of Chip_Ref1, compared to Chip_Ref2 where the variance was more accurately estimated.

This indicates that the bias is mainly due to a form of the “winner’s curse” or “Beavis effect”. That is, a proportion of the selected putative QTN from Ref1 were estimated to have a larger effect than the real effect, and when Chip_Ref1 was used for genomic prediction in the same Ref1 set, these effects are again overestimated. In BayesR the bias was more serious than GBLUP possibly because the BayesR mixture model allows for some large QTN effects, while GBLUP assumes all SNP effects are sampled from a single distribution so that larger effects are regressed more towards the mean. This phenomenon of bias was also reported by Veerkamp *et al.* (2016) using dairy cattle data. However, in the studies by Brøndum *et al.* (2015) and van den Berg *et al.* (2016) the bias was less apparent, most likely because their putative QTN discovery population did not exactly overlap with the genomic prediction reference populations. Wiggans *et al.* (2016) did not test for bias in their study. It might be expected that bias and reduced accuracy may be exacerbated if a GWAS is used to select the top putative variants because the Beavis effect is likely to be more pronounced with SNP effects fitted as fixed effects.

In conclusion, it is important to recognise the pitfalls of pre-selecting subsets of SNP for genomic prediction and to take steps to mitigate them, such as using independent reference populations for QTN discovery and genomic prediction. A potential alternative which does not require two independent populations is a new analytical approach (van den Berg *et al* 2017 - these proceedings) derived from a hybrid method of Expectation-Maximisation with BayesR (HyB_BR) developed by Wang *et al.* (2016).

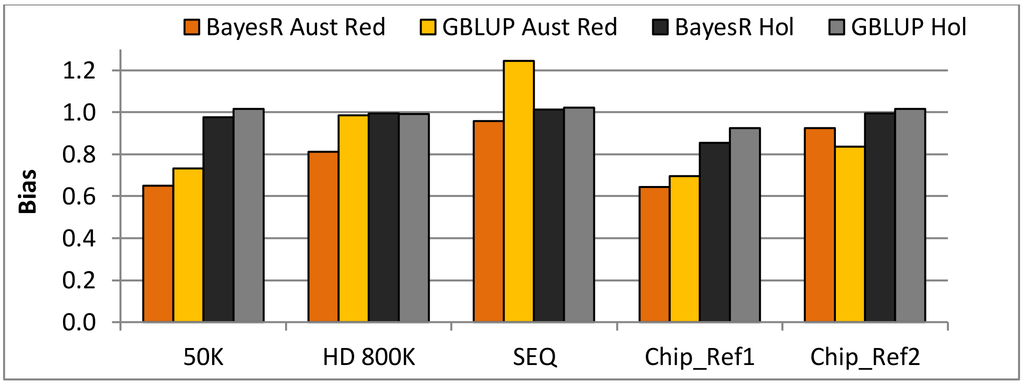

Figure 2. Bias of genomic prediction equations trained in Ref1 and validated in Holsteins or Australian Reds using a range of SNP genotypes. SEQ represents ~1million sequence variants and includes the SNP chosen as surrogate QTN. Chip_Ref1 is a custom chip of 50K + 500 top putative QTN discovered in the same Ref1, while Chip_Ref2 is a custom chip with 50K + 500 top putative QTN discovered independently in Ref2.

ACKNOWLEDGEMENTS

We acknowledge DataGene (and CRV Netherlands for providing access to some of the genotypes used in this study. We acknowledge our partners in the 1000 Bull Genomes Project for access to the reference genomes.

REFERENCES

Brøndum R.F., Su G., Janss L., et all. (2015) *J. Dairy Sci.* **98**: 4107.

Calus M.P.L., Bouwman A.C., Schrooten C. and Veerkamp R.F. (2016) *Gen. Sel. Evol.* **48**: 49.

MacLeod I.M., Bowman P.J., Vander Jagt C.J., et all. (2016) *BMC Genomics* **17**: 144.

van Binsbergen R., Calus M.P.L., Bink M.C.A.M., van Eeuwijk F.A., Schrooten C. and Veerkamp R.F. (2015) *Gen. Sel. Evol.* **47**: 71.

van den Berg I., Boichard D. and Lund M.S. (2016) *Gen. Sel. Evol.* **48**: 83.

Veerkamp R.F., Bouwman A.C., Schrooten C. and Calus M.P.L. (2016) *Gen. Sel. Evol.* **48**: 95.

Wang T., Chen Y.-P.P., Bowman P.J., Goddard M.E. and Hayes B.J. (2016) *BMC Genomics* **17**: 1.

Wiggans G.R., Cooper T.A., VanRaden P.M., Van Tassell C.P., Bickhart D.M. and Sonstegard T.S. (2016) *J. Dairy Sci.* **99**: 4504.