ASSESSING THE VALUE OF WHOLE GENOME SEQUENCE DATA IN SELECTING FOR AGE AT PUBERTY IN TROPICALLY ADAPTED BEEF HEIFERS

C. Warburton and B.J. Hayes

Queensland Alliance for Agriculture and Food Innovation, The University of Queensland,
St Lucia QLD, 4067 Australia

SUMMARY

Age at puberty (AP) has been shown to be heritable in tropically adapted beef heifers, and is associated with lifetime productivity, but it is a difficult and expensive trait to measure. This study investigated whether whole genome sequence (WGS) genome wide association study (GWAS) results could be used to improve the accuracy of selection for AP by using various methods of SNP inclusion and different densities of SNP panels. These results suggest that the most benefit of WGS SNP inclusion would be made in lower density marker panels, with the 6K plus WGS analyses having prediction accuracies equivalent to the 50K base analysis. The ability to use a less expensive, lower density marker panel to make selection decisions will have a financial benefit to producers and warrants further investigation. Further research is required to determine the best technique to select WGS SNP and the most appropriate method to include these SNP into prediction models.

INTRODUCTION

Age at puberty (AP) has been shown to be moderately heritable in tropically adapted beef populations and is favourably correlated to female lifetime reproductive capacity (Johnston *et al.* 2009; Zhang *et al.* 2013; Johnston *et al.* 2014; Farah *et al.* 2016). AP has also been shown to be heritable using genomic information, however, the accuracy of selection using these methods has been low (Zhang *et al.* 2013; Engle *et al.* 2019; Hayes *et al.* 2019).

One possibility for improving accuracy of genomic predictions is to use (imputed) whole genome sequence (WGS). To date, the use of WGS data in genomic predictions within livestock species has shown modest improvements in selection accuracy (0%-5%), and there is much interest in developing novel techniques to best utilise this data (Raymond *et al.* 2018). The aim of this study was to investigate methods to incorporate WGS data into the genomic prediction for AP in a multi-breed population of tropically adapted beef heifers, to improve the accuracy of selection.

MATERIALS AND METHODS

Animals and Phenotypes. Fertility records used in this study were obtained from two research herds, the Northern Breeding Project research herds from the Cooperative Research Centre for Beef Genetic Technologies (Beef CRC) and the Queensland Smart Futures (SMF) population.

Briefly, 868 Brahman heifers and 960 Tropical Composite heifers with both a phenotype for AP and genotype data were obtained from the Beef CRC. In this study, AP was defined as age, in days, at first *corpus luteum*, obtained by ultrasound scanning heifers every 4 to 6 weeks (Johnston *et al.* 2009). Detailed herd structure, management and data recording have been outlined in Johnston *et al.* (2009).

A total of 3,695 reproductive maturity scores (a proxy trait for AP; measured at 600 days by ultrasound and is a 0 to 5 score) were obtained from the SMF database on heifers from 3 breeds, Brahman, Santa Gertrudis and Droughtmaster (Burns *et al.* 2016). Full information on herd structure, management and data recording can be found in Burns *et al.* (2016).

The SMF results presented in this paper have been analysed across breeds to determine if multi-breed genomic predictions could be viable for use in industry data. However, it must be noted that the genomic estimated breeding values (GEBV's) shown in these results are not true multi-breed GEBV's as heifers of each breed were managed separately and there were no mixed breed cohorts analysed in this data.

Genotypes. Beef CRC heifers were genotyped with the BovineSNP50 BeadChip (Illumina, SanDiego, CA) and SMF heifers were genotyped with the 24,121 SNP from the Geneseek GGP-LD array. Full details on genotype quality control are described in Hayes *et al.* (2019). Genotypes were imputed up to 728,785 SNP (Bovine HD array) using the Fimpute software (Sargolzaei *et al.* 2014), and a panel of 1500 cattle of relevant breeds genotyped for the Bovine HD array. All genotypes were then imputed to 23 million whole genome sequence variant genotypes using the 1000 bull genomes Run6 data base (Hayes *et al.* 2019) using Eagle phasing and Minimac3 for imputation.

Statistical analysis. Three datasets, Brahman (Beef CRC), Tropical Composite (Beef CRC) and SMF (Brahman, Santa Gertrudis and Droughtmaster) were used in these analyses. The analysis proceeded in two steps: 1) Identify SNP associated with AP in the imputed sequence data by within breed GWAS analysis in the Beef CRC animals, then 2) Test the accuracy of genomic predictions when these SNPs are added to base SNP panels in the SMF data.

The final models for each analysis included contemporary group fitted as a covariate, which was defined as herd, year and season in the SMF dataset. In the SMF dataset age at AP measurement was also included as a covariate. In the Brahman analysis, age of dam was fitted as a covariate and in the Tropical Composite analyses zebu percentage was fitted as a continuous covariate. Animal was fitted as a random effect in all models.

Two strategies were used to identify SNPs associated with AP in the GWAS:

- TOP GWAS (SNP significance threshold 5.0e-06) - all SNP from the WGS GWAS that met the significance threshold from either breed were included in each analysis.
- TOP META (SNP significance threshold 5.0e-07) - Meta-analyses were conducted on the output from the WGS GWAS of the combined Brahman and Tropical Composite populations using the program Metal (Willer *et al.* 2010) and the SNP that met the significance threshold were included in each analysis.

The numbers of significant SNP from the WGS data for each analysis and each SNP selection strategy are shown in Table 1.

Genomic predictions in the SMF data were conducted using 3 different density of base SNP panels, 6K (BovineLD array), 50K (BovineSNP50 BeadChip) and 800K (BovineHD array). A GBLUP approach was used. Genome-wide complex trait analysis (GCTA) was used to construct genomic relationship matrices (GRM) and perform genomic predictions for each of the datasets for each SNP density, see Yang *et al.* (2011) for more detail.

Significant, unique (not already included in base marker panels) SNP from the sequence GWAS were incorporated into each analysis using one of two methods; first, by adding the significant WGS SNP into the GRM for each analysis (6K plus WGS SNP, 50K plus WGS SNP or 800K plus WGS SNP) or secondly, by using a multi GRM method where the base GRM remained the same but a second GRM, with only the WGS SNP, was added and analysed simultaneously. Any significant WGS SNP that were already included on marker panels were excluded from the WGS GRM but remained in the base GRM in the MGRM analyses. The GEBV from each GRM in the MGRM analyses were added together to calculate total GEBV which was used to calculate prediction accuracy (6K MGRM, 50K MGRM or 800K MGRM).

Five way cross validation within the SMF data set was used to determine the accuracy of prediction of GEBV, where each dataset was randomly split five times and four fifths of the data (reference) was used to predict the GEBV of the last fifth (validation) and the validation animals were then used to calculate the correlation between their predicted GEBV and phenotype adjusted for the model fixed effects. The prediction accuracy was the correlation of the GEBV and the phenotype divided by the square root of the heritability of AP in the SMF 800K analysis, h^2=0.196.

RESULTS AND DISCUSSION

Results in Table 1 show that accuracy was improved more by using a higher density SNP panel than through the addition of WGS SNP in the TOP GWAS analyses. One reason for this may be due to the high level of SNP redundancy that may be occurring with this SNP selection strategy. Of the 165/172 SNP used from the WGS data in the TOP GWAS analysis, a large proportion of SNP occurred on just 3 chromosomes (results not shown), chromosome 1 n=44, chromosome 14 n=77 and chromosome 21 n=18, total number of significant SNP on these 3 chromosomes is 139. There is a probability that a number of these SNP are in close proximity to a single SNP of large effect and, due to linkage disequilibrium, these SNP may appear significant in a GWAS due to this association. Therefore, the actual number of effective SNP that are being used for selection in the TOP GWAS analysis may be lower than the 165/172 shown, which may explain the limited improvement in accuracy seen in Table 1.

Table 1. Prediction Accuracy for TOP GWAS and TOP META analyses in SMF data

	TOP GWAS		TOP META	
Analysis	Prediction accuracy *(s.e)*	No. sig. WGS SNP	Prediction accuracy *(s.e)*	No. sig. WGS SNP
6K	0.36 *(0.04)*		0.36 *(0.04)*	
6K plus WGS SNP	0.37 *(0.05)*	172	0.40 *(0.05)*	1591
6K MGRM	0.37 *(0.04)*	172	0.40 *(0.05)*	1591
50K	0.41 *(0.05)*		0.41 *(0.05)*	
50K plus WGS SNP	0.41 *(0.05)*	172	0.42 *(0.05)*	1587
50K MGRM	0.41 *(0.05)*	172	0.43 *(0.06)*	1587
800K	0.42 *(0.05)*		0.42 *(0.05)*	
800K plus WGS SNP	0.42 *(0.05)*	165	0.42 *(0.05)*	1502
800K MGRM	0.42 *(0.05)*	165	0.44 *(0.05)*	1502

The prediction accuracy of GEBV for TOP META analyses were also improved through the use of higher density SNP panels. In contrast to the TOP GWAS results, the addition of the significant WGS TOP META SNP did result in small improvements in prediction accuracy within each of the analyses, although the improvement is not significant. The inclusion of WGS META SNP into the 6K analysis improved the prediction accuracy of this analysis so that it became equivalent to the 50K analysis. The 6K marker panel is more cost effective for producers than the higher density panels, therefore, if equivalent prediction accuracies can be made from the 6K panel with the use of WGS SNP the financial benefit to producers would be significant.

It is evident that there are many more significant WGS SNP being used in the TOP META analysis, in comparison to the TOP GWAS analysis, which may explain the small improvement in accuracy. Similar to the TOP GWAS results, a large proportion of SNP discovered in the TOP META analysis existed on a single chromosome, 14 (results not shown), n=~1,400 (depending upon the analysis). More research needs to be done to determine the most effective way to select WGS SNP and reduce this potential redundancy.

The MGRM analyses in the TOP META SNP selection strategy resulted in slight improvements in prediction accuracy (though not significant), in comparison to the single GRM analyses, in the 50K and 800K analyses. In the MGRM analysis the WGS SNP are being fitted in their own GRM and, as a result, their effect is less regressed. As these SNP have been selected for having a significant effect upon the AP phenotype from a meta-analysis, it can be argued that fitting these SNP into a single, large GRM may regress their effect by too great an extent. More research is required.

CONCLUSIONS

While the results presented in this paper are not conclusive, there is an indication to suggest that improved methods of WGS SNP selection may be used to improve GEBV prediction accuracy particularly for the less dense marker panels. The inclusion of 1,591 WGS META SNP into the 6K analysis was able to improve the prediction accuracy for puberty to a similar level as the 50K base analysis, which would be a much more cost-effective genotyping solution for producers.

Further research is warranted into appropriate methods to select WGS SNP that are able to explain variation in the AP trait in multi-breed tropically adapted beef populations and the best way to incorporate these SNP into future genomic analysis. More AP phenotypes will be required to improve the accurate detection of WGS SNP that can explain variation in AP across a number of tropically adapted breeds.

ACKNOWLEDGEMENTS

We gratefully acknowledge the Beef CRC and the scientists and technicians who pioneered the traits used in this paper, and the huge effort that went into designing and conducting those experiments. We also gratefully acknowledge the contribution of Dr Brian Burns, Nicholas Corbet, Jack Allen, Alan Laing and Geoffry Fordyce for the SMF data set used in this paper.

REFERENCES

Burns, B.M., Corbet N.J., Allen J.M., A. Laing and M. T. Sullivan (2016).Queensland Government Smart Futures Research Partnerships Program (2012-2015).

Engle, B. N., N. J. Corbet, Allen J.M., Laing A.R., Fordyce G, McGowan M.R., Burns B.M., Lyons R.E. and Hayes B.J. (2019) *J. Anim. Sci.* **97**: 90.

Farah M. M., Swan A.A., Fortes M.R.S, Fonseca R., Moore S.S. and Kelly M.J. (2016) *Anim. Genet.* **47**: 3.

Hayes B. J., Corbet N.J., Allen J.M., Laing A.R., Fordyce G., Lyons R., McGowan M.R. and Burns B.M. (2019) *J. Anim. Sci.* **97**: 55.

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

Johnston D., Barwick S.A., Corbet N., Fordyce G., Holroyd R.G., Williams P. and Burrow H. M. (2009) *Anim. Prod. Sci.* **49**: 399.

Johnston D., Barwick S.A., Fordyce G., Holroyd R.G., Williams P., Corbet N. and Grant T. (2014). *Anim. Prod. Sci.* **54**: 1.

Raymond B.,A. Bouwman C., Schrooten C., Houwing-Duistermaat J. and Veerkamp R. F. (2018) *Genet. Sel. Evol.* **50**: 27.

Sargolzaei, M., Chesnais J. and Schenkel F. (2014) *BMC Genomics* **15**: 478.

Willer C.J., Li Y. and Abecasis G. R. (2010) *Bioinformatics* **26**: 2190.

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

Zhang Y.D., Johnston D.J., Bolormaa S., Hawken R.J. and Tier B. (2013) *Anim. Prod. Sci.* **54**: 16.