GENOMIC PREDICTION USING IMPUTED WHOLE-GENOME SEQUENCE IN AUSTRALIAN ANGUS CATTLE

N. Kamprasert[1], H. Aliloo[1], J. van der Werf[1], C. Duff[2] and S. Clark[1]

[1] University of New England, Armidale, NSW, 2351 Australia
[2]Angus Australia, Armidale, NSW, 2350 Australia

SUMMARY

Using whole-genome sequence data in genomic prediction is expected to improve the predictive ability since the whole genome sequence may contain causal variants. This study aimed to compare the accuracy of genomic prediction with three densities of genotypes, 50k, high- density and whole-genome sequence. The genomic prediction was performed to estimate breeding values for selected growth and carcass traits in Australian Angus beef cattle. Genotype imputation was conducted to retrieve genotypes at high-density and whole-genome sequence level. The dataset was split into testing and reference group to compare the accuracy of breeding values obtained from different genotype densities and for animals with different degrees of relatedness to the reference. The prediction accuracies were similar across three different genotype densities for the traits studied. We found no substantial improvement in genomic prediction accuracy using the whole-genome sequence data in this study.

INTRODUCTION

Genome-based evaluations, commonly known as genomic prediction, have become a standard approach for estimating livestock breeding values. Genomic prediction can improve the rate of response to selection by shortening generation interval and gaining more accuracy in predicting breeding value, especially for young animals and difficult-to-measure traits. The accuracy of the genomic prediction depends on two major factors; the number of DNA-tested animals recorded for the objective trait and the number of DNA markers used in genotyping. Current genomic evaluations use standardised genotyping arrays ranging from 10k to 700k in density, with 50k being the most common platform (Goddard *et al.* 2011). The advent of next-generation sequencing technologies has made it possible to obtain whole-genome sequence data at a reasonable price and such data could be used in routine genetic evaluations. Moreover, genotype imputation is a common practice to obtain whole-genome sequence with a reliable accuracy, for animals genotyped with lower densities.

Whole-genomic sequence is expected to improve the accuracy of genomic prediction since it should include actual causal variants in the data instead of depending on the association between the QTLs and markers (Meuwissen *et al.* 2016). The objective of the present study was to examine the benefit of the sequence data for genomic prediction in Australian Angus beef cattle. Different genetic marker densities, including medium-density 50k, high-density 700k and whole genome sequence were used to examine the potential improvement in prediction ability when increasing the marker density for 3 economically important traits in Australian Angus cattle.

MATERIALS AND METHODS

Animal and data. Data was obtained from the Angus Australia database. The dataset analysed was for animals born between 2013 and 2022. Animals were measured for yearling weight (400dWT), final weight (600dWT) and carcass intramuscular fat (CIMF) (Table 1.). Contemporary groups (CG) were formed according to BREEDPLAN procedures (Graser *et al.* 2005) by concatenating herd, year of birth, sex, birth type, management group defined by breeders and measurement date. The CGs were subdivided by age at measurement with slices of 45 days for the growth traits and slices of 60 days for CIMF. Genotypes for animals were also received from Angus

Australia. Medium-density genotype data (50k) were from the previous study by Aliloo and Clark (2021). A total of 1,076 animals were genotyped with 700k genotype array (HD). Genotype data contained only bi-allelic SNPs located on the autosome.

Genotype imputation. To obtain the whole-genome sequence, genotype imputation was performed. Whole-genome sequence data from 440 Angus bulls from the 1000 bull genome project (Hayes and Daetwyler 2019) were used as a reference for the imputation. The 50k genotype samples were imputed to the whole-genome sequence (WGS) level with a stepwise genotype imputation, from 50k to HD, then to WGS. The genotype imputation was performed with Minimac4 (Das *et al.* 2016) and Eagle (Loh *et al.* 2016) was used for pre-phasing with default parameters. The imputation reference panel was a combination of samples with HD and reduced genotypes from the WGS. The imputation accuracy relied on Miminac4 internal quality metric (Rsq). Post-imputation quality control was applied to the imputed genotypes. Quality control filtered out those SNPs with Miminac4 Rsq < 0.30 and minor allele frequency (MAF) < 0.05. This resulted in 44,827, 522,192, and 7,899,466 SNPs for 50k, HD and WGS, respectively, in the final genotype dataset.

Table 1. Descriptive statistics for growth traits and a carcass trait

	N	**Mean**	**Min.**	**Max.**	**SD.**	**age**	**age mean**
400-day weight, kg	56,058	398.46	235.00	622.00	67.82	301 to 500	400.29
600-day weight, kg	23,705	521.84	339.00	814.00	97.89	501 to 700	574.67
Carcass Intramuscular Fat, %	4,074	9.76	3.00	20.50	3.65	504 to 990	722.24

Statistical Analysis. Possible systematic effects were tested for their significance in the model. The effects tested were CG and a linear and quadratic covariate of age at measurement and dam age. The effects with p-value <0.05 was kept in the final model. Due to a large difference in sample size between growth and carcass traits, they were examined differently. To assess prediction accuracy, a 10-fold cross-validation was conducted using the whole dataset for CIMF. While, for the growth traits, the analysis imitated a forward prediction by splitting animals into a reference and a testing group based on their year of birth. The last two years of the data was used as the testing set and other samples were included in the reference group. Individuals in the testing set were grouped according to the level of their relatedness with the reference set by a relationship value, which extracted from a genomic relationship matrix (GRM). Then, samples in each subgroup were randomly assigned into 10 groups for cross-validation. A univariate animal model using the full dataset with 50k genotype density was used to generate phenotypes corrected for all estimated fixed-effect coefficients. GRMs with three different genotype densities were constructed based on Yang *et al.* (2011) using GCTA software. The top-30 relationship values were extracted from off-diagonal elements of the GRM using 50k and then averaged (Clark *et al.* 2012). Observed phenotypes of the testing samples were masked and genomic estimated breeding values were obtained from analyses based on 50k, HD and WGS genotypes. The genomic prediction was performed using the GBLUP approach with a univariate animal model using MTG2 (Lee and Van der Werf 2016). The accuracy of genomic prediction was calculated as the Pearson correlation coefficient between the corrected phenotypes and GEBVs of the testing group divided by the square root of the trait heritability obtained from a 50k-based analysis. The accuracies with the standard error were expressed as an average value from the cross-validation. The accuracy of genomic predictions was compared between three densities of genotypes, and was reported from the testing group and the subgroups according to the degree of relatedness.

RESULTS AND DISCUSSION

The accuracies of genomic prediction for the studied traits are presented in Table 2. For both growth traits, the prediction accuracies were similar for the three genotype densities. Although there was no significant difference, the HD density had the highest accuracy with values of 0.683 (0.017) and 0.630 (0.016) for 400dWT and 600dWT, respectively. The lowest accuracy for both traits was from the WGS, given 0.675 (0.016) for 400dWT and 0.621 (0.014) for 600dWT. The accuracy marginally increased from 50k to HD, then slightly decreased from HD to WGS. Similarly, there was no difference in the prediction accuracies for CIMF. The highest accuracy was 0.643 (0.027) retrieved from HD but there was not significantly different in a comparison. Our results agreed with previous studies showing that using WGS did not significantly improve the accuracy of genomic prediction (Raymond *et al.* 2018; Bedhane *et al.* 2021).

Table 2. Prediction accuracy[1,2] with three different genotype densities by testing group and by relatedness subgroups, and trait heritability

	n	**50k**	**HD**	**WGS**
400-day weight, kg				
testing group	17,942	0.677 (0.016)	0.683 (0.017)	0.675 (0.016)
medium-related	10,230	0.656 (0.020)	0.659 (0.021)	0.650 (0.022)
high-related	7,712	0.711 (0.019)	0.721 (0.020)	0.714 (0.020)
h^2		0.246 (0.007)		
600-day weight, kg				
testing group	5,117	0.627 (0.014)	0.630 (0.016)	0.621 (0.014)
medium-related	3,259	0.611 (0.013)	0.615 (0.012)	0.608 (0.011)
high-related	1,858	0.659 (0.032)	0.660 (0.036)	0.648 (0.035)
h^2		0.338 (0.001)		
Carcass Intramuscular Fat, %				
testing group		0.639 (0.024)	0.643 (0.027)	0.637 (0.027)
h^2		0.464 (0.027)		

[1] Prediction accuracy with standard error was obtained from the 10-fold cross-validation.
[2] There was no significant difference in a comparison (p-value <0.01).

Prediction accuracy by relatedness group. Different number of top relationship values were tested to define strength of relatedness between testing samples and reference set. The top-30 average was found to clearly split testing set into two groups (Figure 1). Then, the testing set was subdivided into two groups, which were medium- and high-related groups, and 0.25 was the threshold point. There were 17,942 and 5,117 animals in the testing group for 400dWT and 600dWT, respectively. For CIMF, a 10-fold cross-validation with the whole dataset was performed.

As expected, the high-related group obtained more accurate predictions compared to the medium-related group (Table 1). The accuracy of relatedness subgroups fluctuated with only a slight change with different genotype density. However, difference in the accuracy was not significant between the genotype densities. The highest accuracy for both the medium- and the high-related group were from HD, and the lowest accuracy was from the WGS. The accuracy by subgroups was similar to the testing group where the accuracy steadily declined as the genotype density increased. There were a small difference and no clear pattern in the prediction accuracy when increasing the genotype density. Lastly, accuracy of genomic prediction is involved by several factors, for instance, trait heritability, size of the reference and relatedness between selection samples and the reference.

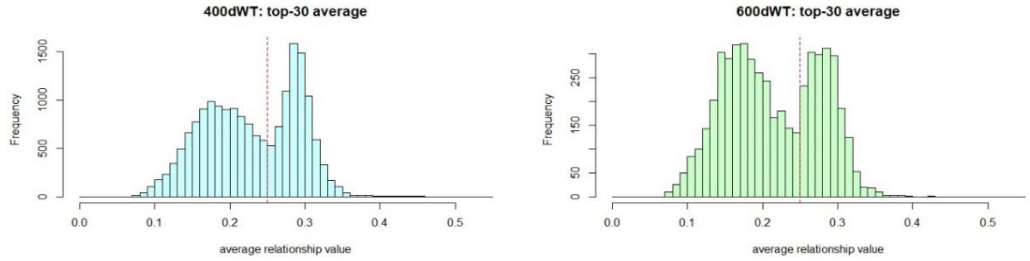

Figure 1. Relationship values for testing animals from 50k genotype by traits with the threshold point

CONCLUSION

This study has investigated the benefit of whole-genome sequence for predicting breeding values for the selected growth and carcass traits in Angus cattle. Although the highest prediction accuracies were retrieved when using the high-density genotype, the difference was not significant compared to the 50k-based prediction. For the traits studied, there was no clear evidence of increased prediction accuracy with denser genotypes, such high-density array and whole-genome sequence.

REFERENCES

Aliloo H. and Clark S.A. (2021) *Animal Production Science* **61:** 1958.

Bedhane M., van der Werf J.H.J., de las Heras-Saldana S., Lim D., Park B., Park M.N., Hee R.S., Clark S. and Fortes M. (2021) *Animal Production Science* **62:** 21.

Clark S.A., Hickey J.M., Daetwyler H.D. and van der Werf J.H.J. (2012) *Genetics Selection Evolution* **44:** 4.

Das S., Forer L., Schönherr S., Sidore C., Locke A.E., Kwong A., Vrieze S.I., Chew E.Y., Levy S., McGue M., Schlessinger D., Stambolian D., Loh P.-R., Iacono W.G., Swaroop A., Scott L.J., Cucca F., Kronenberg F., Boehnke M., Abecasis G.R. and Fuchsberger C. (2016) *Nature Genetics* **48:** 1284.

Goddard M.E., Hayes B.J. and Meuwissen T.H.E. (2011) *J. Anim. Breed. Genet.* **128:** 409.

Graser H.-U., Tier B., Johnston D.J. and Barwick S.A. (2005) *Aust. J. Exp. Agric.* **45:** 913.

Hayes B.J., Daetwyler H.D. (2019) *Annual Review of Animal Biosciences* **7:** 89.

Lee S.H. and van der Werf J.H.J. (2016) *Bioinformatics* **32:** 1420.

Loh P.-R., Danecek P., Palamara P.F., Fuchsberger C., A Reshef Y., K Finucane H., Schoenherr S., Forer L., McCarthy S., Abecasis G.R., Durbin R., L and Price A. (2016) *Nature Genetics* **48:** 1443.

Meuwissen T., Hayes B., Goddard M. (2016) *Animal Frontiers* **6:** 6.

Raymond B., Bouwman A.C., Schrooten C., Houwing-Duistermaat J. and Veerkamp R.F. (2018) *Genetics Selection Evolution* **50:** 1.

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