THE ACCURACY OF GENOTYPE IMPUTATION IN SELECTED SOUTH AFRICAN SHEEP BREEDS FROM AUSTRALIAN REFERENCE PANELS

C.L. Nel[1], K.P. Gore[2], A.A. Swan[2], S.W.P. Cloete[1,3], J.H.J. van der Werf [4] and K. Dzama[1]

[1]Department of Animal Science, University of Stellenbosch, Stellenbosch, WC, 7602 South Africa
[2]Animal Genetics & Breeding Unit*, University of New England, Armidale, NSW, 2351 Australia
[3]Directorate Animal Sciences, Western Cape Department of Agriculture, Elsenburg, 7607 South Africa
[4]School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351 Australia

SUMMARY

The cost of genotyping is becoming increasingly affordable but remains an influential factor for determining the SNP-density at which genotyping can proceed. Compared to Australian breeding programs, the South African wool sheep industry represents parallel objectives within similar environments but presently lacks the necessary infrastructure to exploit modern technologies such as genomic selection. The aim of the study was to determine the feasibility of across country imputation as an alternative to high density genotyping on a local basis. Following imputation from a 15k to 50k density, mean accuracy levels of 0.87 and 0.85 were observed in the Merino and Dohne Merino breeds, while the highest levels of accuracy of 0.88 and 0.90 was observed in the Dorper and White Dorper breeds, respectively. The extent of genetic relationships was considered amongst the key factors that limit the ability to impute at an accuracy above 90%, but the observed results suggest that across country imputation could remain useful. Imputation from reference panels genotyped at densities higher than 50k and research into across country prediction is recommended.

INTRODUCTION

Genomic prediction and Genome Wide Association Studies (GWAS) depend on the size of the reference population as well as the density at which informative individuals were genotyped. Even though medium and high density genotyping options are becoming more affordable, cost remains a restricting criterion for the choice of a genotyping platform. Economic restrictions are likely to be more severe within a developing infrastructure as is currently experienced in South Africa (Van Marle-Köster and Visser 2018). There could be potential to exploit similarities in South African and Australian *ovine* breeds and environments through the compilation of genotypic resources. Imputation of un-typed markers of animals genotyped at a lower density has proven a reliable and affordable alternative to widespread genotyping on high density platforms (Browning and Browning 2007; Berry and Kearney 2011; Hickey *et al.* 2011; Huang *et al.* 2012; Moghaddar *et al.* 2015). The objective of this study is thus to investigate the potential of across country imputation of South African datasets from Australian reference populations from low (15k) to medium (50k) densities.

MATERIALS AND METHODS

Data Structure and Distribution. The South African sample set was selected from multiple breeds within respective resource flocks (Schoeman *et al.* 2010) as well as a smaller proportion of animals originating from the industry sector. Genotyping of the South African (SA) sample set was performed with the OvineSNP50 (Illumina Inc., CA, USA) beadchip at GeneSeek Inc. (Lincoln, NE,

* A joint venture of NSW Department of Primary Industries and the University of New England

USA) and subjected to quality control measures (> 0.25 GenCall score, > 0.5 GenTrain score, > 0.01 MAF, > 0.95 call rate, > 0.95 sample call rate). Following imputation of randomly missing SNPs, 986 samples with 50095 SNPs remained available for further analysis. Animals were grouped by breed type, namely Merino (552), Dohne Merino (60), Dorper (59), South African Mutton Merino (57), Dormer (42), Meatmaster (39) and White Dorper (27) while the hardy native breeds Damara (30), Pedi (29) and Afrikaner-type (13) animals were grouped together as 'Indigenous' (72). The Australian reference set constituted a database of ~ 84 000 samples from multiple breeds that serve as respective reference populations in genomic prediction programs. The major proportions of the dataset were classified as Merino, maternal (Border Leicester and Coopworth) and terminal (Poll Dorset and White Suffolk) groups. The same OvineSNP50 genotyping platform was used in generating the Australian database and 48 599 SNPs were available for analysis following quality control.

Design. All 986 SA samples were subset to ~ 15k SNPs using Illumina map information to simulate a commercial 15k beadchip. Analysis proceeded by the subsequent imputation back up to the 50k density using an Australian reference. The accuracy of imputation was evaluated by Pearson correlation coefficients between the imputed and observed SNP genotypes. To reduce computation time and increase accuracy (Moghaddar *et al.* 2015), the Australian reference set was screened by assigning an animal in the sample set with the top 50 highest values of animals in the reference set according to a genomic relationship matrix (GRM) that included all the animals in the study. Thus, animals from the reference set not meeting this criterion for any of the animals in the sample set were not used for imputation.

Software. Genotype imputation was performed using FIMPUTE (V2.2) (Sargolzaei *et al.* 2014). The program assumes a level of relatedness between all individuals and phases reference sets with overlapping sliding windows that is shrunk in proceeding increments with each chromosome sweep. The initial larger window sizes aim to capture the long-range haplotypes expected from highly related individuals, while the subsequent sweeps aim to capture relationships between more distant individuals. The inclusion of pedigree information is an optional addition to FIMPUTE, but it was not supplied in the current analysis. Summary statistics and visual analyses were performed in R (R Core Team 2016, Vienna, Austria).

RESULTS AND DISCUSSION

The accuracy of imputation varied considerably both between and within breeds. The accuracy of imputation for indigenous breed group was very low (mean = 0.68) and is not represented in subsequent figures and tables. A low accuracy is to be expected considering their heterogeneous nature and poor representation within the Australian reference set. Moreover, concerns have been raised surrounding an underrepresentation of indigenous breeds in the design of commercial bead chips (Sandenbergh *et al.* 2016). Table 1 shows the summary statistics for imputation accuracy (correlation coefficients) for the remaining breed groups.

The accuracy for Merino samples was moderate, as Pearson's correlations ranged from 0.82 to 0.90. This is considerably lower than correlation coefficients of 0.93 to 0.96 reported by Moghaddar *et al.* (2015) for 1,000 purebred Merinos imputed from smaller proportions of the same reference set. Hayes *et al.* (2012) reported considerably lower values of accuracy (71%) for imputing Merino samples from 5k to 50k densities, but with a reference set confined to ~ 5000 animals. Moderate accuracies were also observed for Dohne Merino and the South African Mutton Merino (SAMM) individuals, while the imputation accuracy for Dormers and Meatmasters were below 0.80.

Table 1. Summary statistics for the imputation accuracy of all South African breed groups in the sample set

	Merino	Dohne Merino	Dorper	SAMM	Dormer	Meat -master	White Dorper
(n)	552	60	59	57	42	39	27
Min.	0.82	0.82	0.85	0.81	0.76	0.72	0.87
1st Quartile	0.86	0.85	0.87	0.83	0.78	0.74	0.89
Mean	0.87	0.85	0.88	0.85	0.79	0.75	0.90
3rd Quartile	0.87	0.86	0.89	0.86	0.79	0.76	0.91
Max.	0.90	0.88	0.90	0.87	0.81	0.78	0.92

The Dorper and White Dorper breeds achieved moderately high to high imputation accuracies. The Dorper originates from South Africa, and it is possible that the animals that represent them in the Australian database have not drifted extensively from the ancestral lines or share relatively recent parental links. Considering the size of the Australian reference set and the large proportion of Merinos included, it could be considered somewhat unexpected that none of the Merino test samples attained an imputation accuracy of above 0.90. However, the number of reference samples available as well as their relatedness to the sample population is considered essential factors in the accurate phasing of haplotypes for the imputation of un-typed markers.

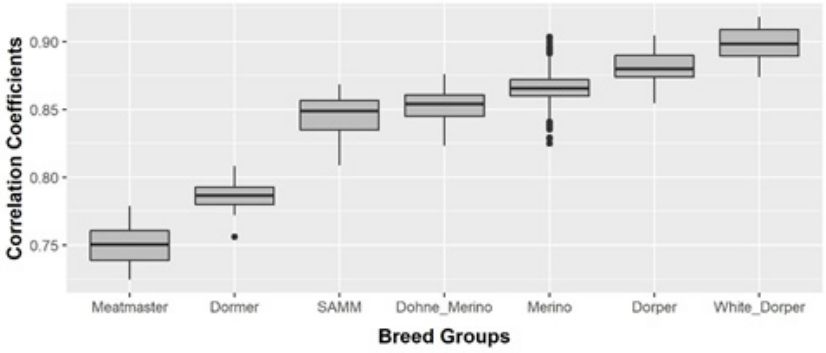

Figure 1. Box plots for the imputation accuracies for all South African breeds in the sample set

Analyses that characterize haplotypes using population linkage disequilibrium (LD) based methods do not utilize a pedigree, but indirectly capture patterns associated with identity by descent (IBD), the accuracy of which is complemented by the indication that there is little benefit in including pedigree data if the reference set is large enough (Browning and Yu 2009; Larmer *et al.* 2014; Moghaddar *et al.* 2015). Hayes *et al.* (2012) proposes that haplotypes are not necessarily shared across breeds and that 50k genotyping platforms do not capture LD to an adequate level for across breed application. With markedly less family linkage, the proportion of genomic regions possibly considered IBD should be markedly smaller when attempting across country imputation. Thus, a similar argument to that proposed by Hayes *et al.* (2012) could be extended to the current results, despite the current study being within breed analysis. It is possible the denser 500k platform could provide improved phasing of the reference set that is less dependent on long range haplotypes and more appropriate for capturing linkage disequilibrium observable over distant populations.

CONCLUSIONS

Genotype imputation of un-typed markers in a population depends on the representation of that population within a reference set. There is little benefit in the addition of genetically dissimilar animals. Across country imputation will likely be limited by a lack of direct genetic links, but moderately high levels of accuracy can still be achieved within breeds. Research into across country genomic prediction for shared breeds is recommended.

REFERENCES

Berry D.P. and Kearney J.F. (2011) *Animal* **5**: 1162.
Browning S.R. and Browning B.L. (2007) *Am. J. Hum. Genet.* **81**: 1084.
Browning B.L. and Yu Z. (2009) *Am. J. Hum. Genet.* **85**:847.
Hayes B.J., Bowman P.J., Daetwyler H.D., Kijas J.W. and van der Werf J.H.J. (2012) *Anim. Genet.* **43**: 72.
Hickey J. M., Kinghorn B. P., Tier B., Wilson J. F., Dunstan N., and van der Werf J. H. J. (2011) *Genet. Sel. Evol.* **43**: 1.
Huang Y., Hickey J.M., Cleveland M.A. and Maltecca C. (2012) *Genet. Sel. Evol.* **44**: 1.
Larmer S.G., Sargolzaei M. and Schenkel F.S. (2014) *J. Dairy Sci.* **97**: 3128.
Moghaddar N., Gore K.P., Daetwyler H.D., Hayes B.J. and van der Werf J. H. J. (2015) *Genet. Sel. Evol.* **47**: 1.
Sandenbergh L., Cloete S., Roodt-Wilding R., Snyman M. A. and Bester-van der Merwe A. E. (2016) *S. Afr. J. Anim. Sci.* **46**: 2011.
Sargolzaei M., Chesnais J.P. and Schenkel F.S. (2014) *BMC Genomics* **15**: 478.
Schoeman S.J., Cloete S.W.P. and Olivier J.J. (2010) *Livest. Sci.* **130**: 70.
Van Marle-Köster E. and Visser C. (2018) *S. Afr. J. Anim. Sci.* **48**: 808.