Contributed paper

THE EFFECTS OF NUMBER OF REFERENCE INDIVIDUALS ON THE ACCURACY OF IMPUTATION FROM LOW AND MEDIUM DENSITIES TO HIGH DENSITY

M.H. Ferdosi[1], N.K. Connors[1] and M. Khansefid[2]

[1]Animal Genetics & Breeding Unit, University of New England, Armidale, NSW, 2351 Australia.
[2] AgriBio Centre for AgriBioscience, Agriculture Victoria, Bundoora, VIC 3083, Australia

SUMMARY

Imputation is a common approach to infer the missing markers for individuals with low marker density (i.e. target population) from a reference population genotyped with higher-density Single-Nucleotide Polymorphism (SNP) panels. Several factors affect the imputation accuracy of untyped, including the number of reference individuals, marker density and population structure. This paper investigates the effects of these factors on the accuracy of imputation by using individuals of a single cattle breed or multiple cattle breeds in the reference population with 600k marker density, as well as assuming the target population was genotyped with low (15k) or medium (30k) marker density. To achieve a within breed imputation accuracy of >90%, we required at least 500 individuals in the reference population when the target population was genotyped with 15k SNP panel. Whereas, if the reference population consisted of a mixture of purebred and multi-breed individuals, the SNP density must be at least 30k in the target population, and there must be more than 900 individuals in the reference population to achieve a similar level of accuracy.

INTRODUCTION

Genotyping thousands of individuals per month for genomic evaluations has become a common practice in livestock industries in many countries to increase the rate of genatic gain. To reduce the costs of genotyping, industry animals are often genotyped with medium-density panels. Previous studies show that imputing genotypes to high-density and sequence variants can increase genomic prediction accuracy and improve genome-wide association power of Quantitative Trait Loci detection (Moghaddar *et al*. 2019; Khansefid *et al*. 2020). Several factors influence the imputation accuracy of untyped SNPs, such as the number of individuals with high-density markers (i.e. reference population size), the density of markers in the reference and target population, and population structure (Browning and Browning 2011; Ferdosi and Connors 2019; Connors and Ferdosi 2020). The population structure in imputation studies generally refers to the genetic relatedness of individuals within and between reference and target populations. In this study, we investigated the effect of genotyping the target population with higher-density markers and increasing the size of the reference population on the imputation accuracy. Genotypes were imputed from varying medium densities to high density (582k), with reference populations varying in number and breed. The size of the reference population was increased by including more individuals of similar breeds to the target population in the "single-breed reference" or including individuals of multiple breeds in the "multi-breed reference".

MATERIALS AND METHODS

Genotypes. Genotypes were extracted using the BREEDPLAN genomic pipeline (Connors *et al.* 2018; Johnston *et al.* 2018). The individuals and SNPs which had missing rates greater than 5% and the SNPs with minimum allele frequency (MAF) lower than 5% were removed. For multi-breed

[1] AGBU is a joint venture of NSW Department of Primary Industries and University of New England

imputation study, the genotypes of 4,458 individuals and 682k SNPs were reduced to 4,363 individuals and 624k SNPs after quality control (QC). The multi-breed populations included Angus (387), Brahman (610), Charolais (730), Hereford (294), Limousin (742), Santa Gertrudis (213), Wagyu (75), Simental (213), Shorthorn (123) and minor breeds (976). For the single-breed imputation study, the relationship between the individuals in the target and reference populations had to be greater than 0.8 (Boerner and Wittenburg 2018). Genotypes of 618 Brahman and 748 Charolais with 682k SNPs were extracted, reducing after QC to 609 Brahmans with 579k SNPs, and 728 Charolais with 582k SNPs.

Reference and target populations. A proportion of individuals with high-density genotypes were selected as a reference population, and the genotypes of the remaining individuals were converted to 15k and 30k densities by masking a random set of SNPs. Hence, in the randomly selected individuals for the target populations, some of the known genotypes were converted to missing genotypes and this procedure was repeated 9 times for each scenario.

In the multi-breed imputation study, the reference populations were consisting of 100, 200, 300, 400, 500, 600, 800, 1000 and 2000 individuals. While in the single-breed imputation study, the reference populations only consisted of 100, 200, 300, 400, 500, and 600 individuals.

Imputation. FImpute Version 2.2 with default parameters (Sargolzaei *et al.* 2014) was used to impute missing genotypes using single or multiple breeds in the reference without exploiting known pedigree information.

Imputation accuracy. Pearson's correlation coefficient between true and imputed genotypes for individuals was calculated to assess the accuracy of imputation in the different scenarios.

RESULTS AND DISCUSSION

Figure 1 shows the correlation coefficients between true and imputed genotypes in different scenarios. In general, increasing the number of individuals in the reference population and increasing the number of SNPs in the target population from 15k to 30k improved the imputation accuracies for all scenarios. These results were expected and in line with the previous reports (Ferdosi and Connors 2019). Using the same breed in the reference and target populations led to higher imputation accuracy compared to including multiple breeds in the reference. For the purebred individuals with 15k SNPs, there should be more than 500 individuals in the reference population to achieve imputation accuracy higher than 0.9, while with 30k SNPs, 200 individuals in the reference population were sufficient to achieve a similar level of accuracy. For multi-breed, the number of individuals in the reference population and the number of SNPs in the target population needed to be higher compared to single-breed, to achieve a correlation higher than 0.9. Imputation accuracy for a few individuals remained low (shown as outliers in Figure 1) in all scenarios probably because some haplotypes in the target population were undetected or incorrectly detected in the reference population even after including more individuals in the reference. For example, for imputing from 30k SNP to high-density by using 2000 individuals in the multi-breed reference, 54 individuals had imputation accuracy less than 0.85 and 51 of those individuals had a relationship to the relevant breed reference population less than 0.8. This indicates that imputation accuracy tends to be lower in multibreed populations compared to purebreds.

The genotypes from industry animals are used in genomic evaluation and GWAS (i.e. finding QTL) for many traits with diverse range of heritabilities. The accuracy of genomic predictions for the traits with high levels of heritability might be just marginally improved by increasing the marker density through imputation. However, imputation could be still useful to increase the power of QTL detection and especially for improving the accuracy of predictions for the traits with low levels of heritability or when the number of animals in the reference population is limited (Moghaddar *et al*. 2019). Moreover, in terms of practicality, it is much easier to use a same set of SNPs (i.e. imputed to high-density) in genomic evaluation of all traits regardless of their heritability.

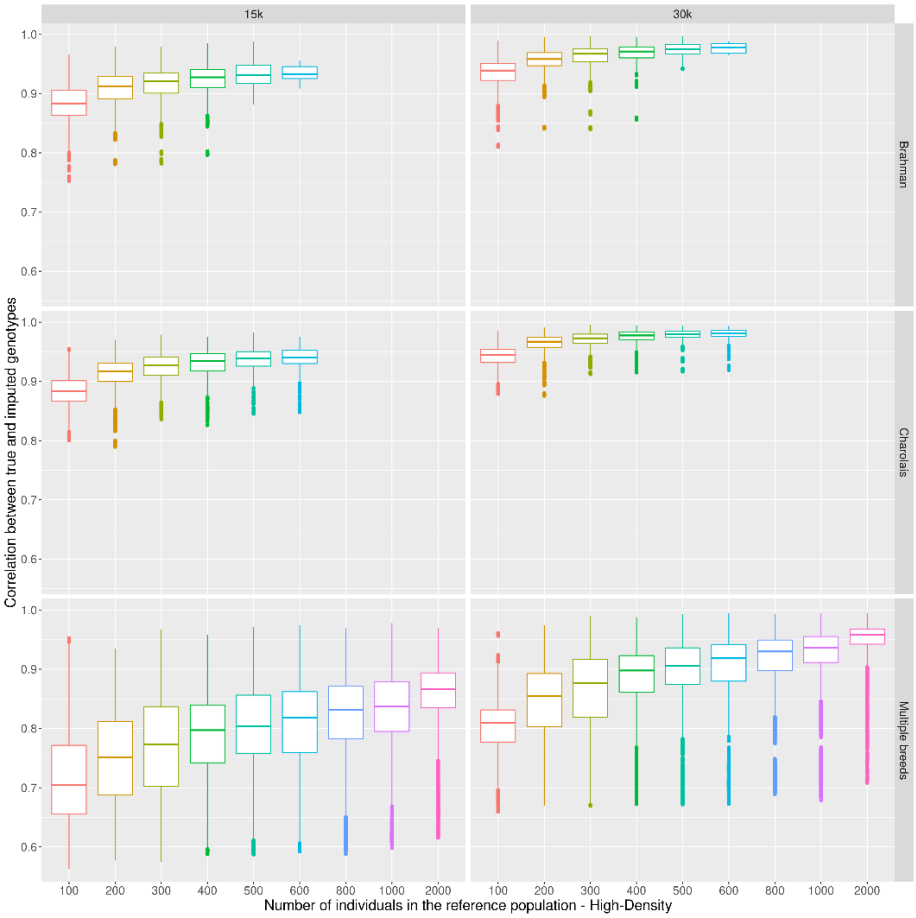

Figure 1. Correlation between true and imputed genotypes for 3 reference populations and 2 marker densities. The boxplots show the correlation coefficients between true and imputed genotypes in different imputation scenarios. The 15k is low density and 30k is medium density panels

CONCLUSIONS

In this study, we explored the effect of the number of SNPs, the number of individuals in the reference and using a single or a multi-breed reference population on the imputation accuracy. The results showed that imputation accuracy was higher when the reference and target populations were of the same breed. In a multi-breed reference population with even a large number of individuals, the imputation accuracy was low, i.e. despite the number of individuals increased in the reference population, the imputation accuracy was lower than purebred scenarios. Increasing the SNP density of the target population to 30k, as well as increasing the number of individuals in the reference population, could improve the imputation accuracy. Algorithms behind the imputation programs are also important and further studies should evaluate how different algorithms affect the imputation accuracies in various scenarios.

ACKNOWLEDGEMENTS

This study was supported by Meat and Livestock Australia project L.GEN.1704. The authors

thank the Australian beef societies and Irish Cattle Breeding Federation for providing the data for this study.

REFERENCES

Boerner V., Wittenburg D. (2018) *Front Genet* **9:**

Browning S.R., Browning B.L. (2011) *Nat Rev Genet* **12:** 703.

Connors N.K., Ferdosi M.H. (2020) *6th International Conference on Quantitative Genetics.* 135.

Ferdosi M., Connors N. (2019) *Proc. Association for the Advancement of Animal Breeding and Genetics.* 480.

Khansefid M., Goddard M.E., Haile-Mariam M., Konstantinov K.V., Schrooten C., de Jong G., Jewell E.G., O'Connor E., Pryce J.E., Daetwyler H.D., MacLeod I.M. (2020) *Front Genet* **11:** 598580.

Moghaddar N., Khansefid M., van der Werf J.H.J., Bolormaa S., Duijvesteijn N., Clark S.A., Swan A.A., Daetwyler H.D., MacLeod I.M. (2019) *Genet Sel Evol* **51:** 72.

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