INVESTIGATION INTO THE EFFECTS OF NUMBER OF SNPS AND NUMBER OF REFERENCE INDIVIDUALS ON IMPUTATION ACCURACY

M.H. Ferdosi and N.K. Connors

Animal Genetics & Breeding Unit*, University of New England, Armidale, NSW, 2351 Australia

SUMMARY

Imputation is used in Australia's national beef genetic evaluation, BREEDPLAN, to resolve missing genomic information in the combined panels of genomic data and clarify relationships through building a genomic relationship matrix (GRM). The accuracy of imputation is dependent on many factors such as the size of reference population and the density of markers. Here we demonstrated an improvement in imputation accuracy of Angus genotypes by increasing the numbers of individuals and the density of markers in the reference population. The results show a considerable increase in imputation accuracy by increasing density of markers up to 15k SNPs.

INTRODUCTION

Imputation of missing genotypes or imputation from low to high density genotypes has become common practice before building the GRM for routine breeding value estimation through single step best linear unbiased prediction (ssBLUP). Imputation can be used to make a common Single Nucleotide Polymorphism (SNP) panel from several varying density SNP panels, enabling all genotypes to be analysed together, and build the GRM (VanRaden 2008). However, the accuracy of imputation can vary and needs to be addressed. Most imputation algorithms firstly phase the genotypes and then impute the missing SNPs. Therefore, factors that affect haplotype inference accuracy also influence imputation accuracy. The main factors affecting imputation accuracy are the sample size of the reference population (high density genotypes) and to some extent, the target population (low density genotypes to be imputed), marker densities, genotyping accuracy, relatedness both between and within reference and target populations, allele frequencies of markers, and population structure (Browning and Browning 2011). In this study we evaluate the effects of number of SNPs and size of reference population on imputation accuracy of Angus genotypes.

MATERIALS AND METHODS

Genomic data. Angus genotypes included in the November 2018 TransTasman Angus BREEDPLAN analysis were used to assess the imputation accuracy, after quality control using the BREEDPLAN genomic pipeline (Connors *et al.* 2017) were used to assess the imputation accuracy. A subset of 11,041 individuals with 32,453 SNPs were selected from 56,369 individuals, where they had less than 5% missing SNPs in loci and individuals had less than 1% missing genotype (call rate greater than 99%). The missing rate in the final dataset was very low (0.2 percent).

Subsets of individuals to assess the imputation accuracy. Genotypes were divided into reference population subsets based on the individual's date of birth, resulting in 7 subsets with quantities of 777, 1,717, 2,945, 4,148, 6,063, 7,947 and 9,766 individuals, respectively (shown coloured in Figure 1). The last subset (subset 8) was not used in the reference population. Each reference subset was removed from the 11,041 total individuals, such that the remaining animals formed the target population. For example, in the first subset there were 777 individuals in the reference population and the remaining 10264 individuals formed the target population; the second subset had 1,717 reference and 9,324 target individuals, and so on. Subsets were cumulative, i.e. the second reference subset with 1,717

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

individuals included the first subset's 777 individuals.

Subsets of SNPs to assess the imputation accuracy. All genotypes originally consisted of 32,453 SNPs. Each reference population mentioned above used all 32,453 SNPs to impute the SNPs of the target populations. The target populations had subsets of SNPS retained as original, with the remaining SNPs masked, to be imputed back. Retained SNPs were selected randomly each time (and sampled only once), with subsets of 500, 700, 1000, 1500, 2000, 3000, 4000, 5000, 7000, 10000, 15000, 20000 and 30000 SNPs. The SNPs that were not in these subsets (masked SNPs) were imputed back and compared with the original SNPs to check the imputation accuracy.

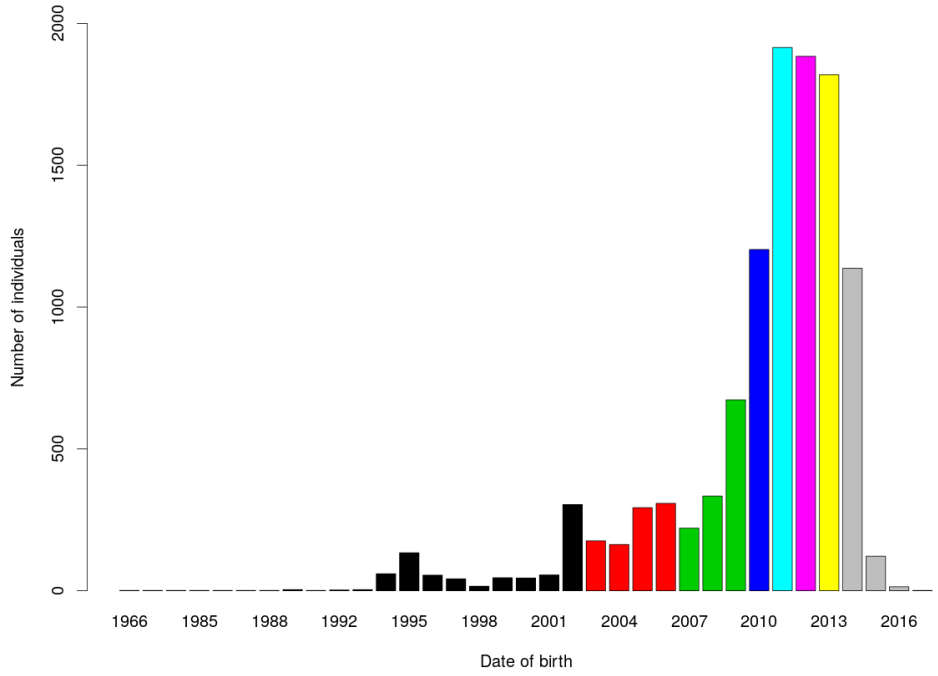

Figure 1. Frequency of number of individuals by their date of birth. Each of the colours represents a different subset used in this study (accumulative, i.e. the second group includes individuals with both red and black colours). In total 8 subsets were used

Imputation. FImpute Version 2.2 and default parameters (Sargolzaei *et al.* 2014) was used to impute missing genotypes without using pedigree information

Imputation Accuracy. The imputation accuracies were calculated for both SNPs and individuals. For individuals, the imputation accuracy was equal to the number of masked SNPs imputed correctly divided by the total number of non-missing SNPs for each individual. A single inaccurately imputed individual can result in multiple changes in relationship between genotyped individuals and reflected in the GRM. As such the minimum imputation accuracy for individuals is used to quantify the effects of changing reference and SNP numbers.

Imputation accuracy of SNPs is calculated by dividing the number of correctly imputed SNPs by the number of SNPs being imputed in that subset. As genomic relationships between individuals depends on all SNPs, the mean imputation accuracy of SNPs was used to quantify effects of changing reference and SNP numbers.

RESULTS AND DISCUSSION

Increasing the number of SNPs and the number of individuals in the reference population resulted in an increase in imputation accuracies of both SNPs and individuals (Figure 2). There was much higher variation in SNP imputation accuracy than that of individual imputation accuracy. The increase in both imputation accuracies was substantial when using up to 15,000 SNPs, at which point the increase in accuracy plateaus. The increasing number of SNPs in each subset decreased the number of SNPs to be imputed, and as such increased potential linkage disequilibrium (LD) between SNPs, even though they were chosen randomly. While Figure 2(B) demonstrates the mean imputation accuracy of SNPs, some SNPs showed much lower accuracy than the mean. For example, some SNPs imputed from the 700 SNPs subset had an imputation accuracy close to 0.1 (data not shown) and some SNPs imputed from the 30,000 SNPs subset had an imputation accuracy close to 0.4. The low imputation accuracy for these SNPs may indicate that these SNPs were not mapped correctly, or they were in regions with high recombination rate (e.g. recombination hotspots). In addition, it may indicate that these SNPs may be in regions where the quality of genotyping is generally low.

The number of individuals in the reference population was also important however with less effect in comparison to the number of SNPs. Figure 2 (A) shows an increase in imputation accuracy for each reference population increasing in individuals. Based on Figure 2 (B) more than 2,000 individuals are required in order to see substantial improvement in the SNPs imputation accuracies. In this study, we reported the imputation accuracy for all the subsets, although there was not considerable difference between subsets. A better approach may be to consider the accuracy for every immediate generation. Only genotype information was considered to evaluate imputation accuracies, however if reliable pedigree is available the imputation accuracy can be increased, especially if less than 15,000 SNPs are used. Here, the subsets of SNPs were selected randomly, rather than selecting SNPs based on their LD or other methods, as well as including SNPs with low minimum allele frequencies (<0.05). Further study is required to identify the effect of pedigree on the imputation accuracy of genotypes, along with SNP selection methods and allele frequency. Importantly, by selecting numbers of individuals based on date of birth, the results may be affected by population structure. Increasing the number of individuals in the reference population also increased relationships between reference and target populations, due to the structure of subset populations. The results showed the accuracy of imputation was affected by density of markers more than the number of animals in the reference population.

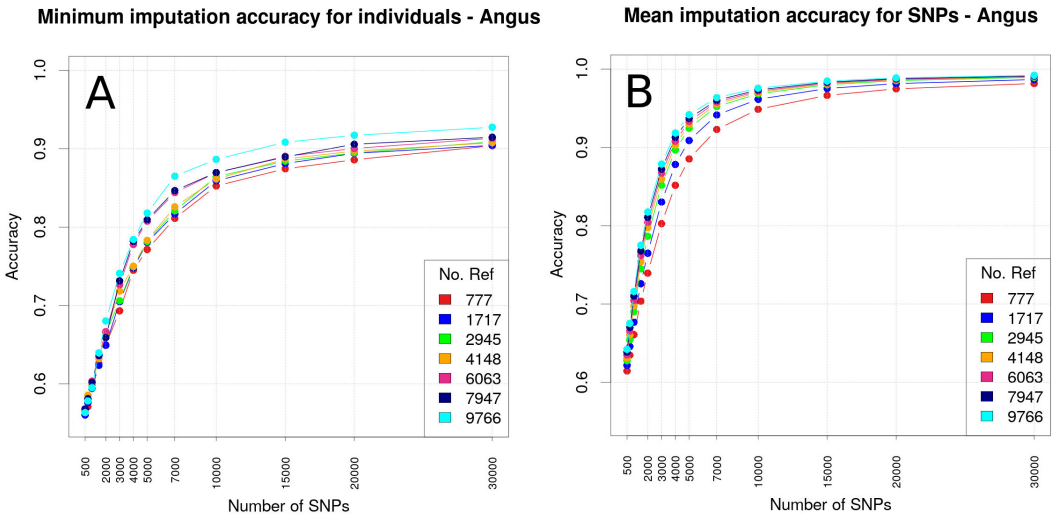

Figure 2. Effect of the number of SNPs on imputation accuracy. (A) Effect of the number of SNPs on the minimum imputation accuracy of individuals. (B) Effect of the number of SNPs on the mean imputation accuracy of SNPs

CONCLUSIONS

In this study we showed how the number of SNPs and the number of individuals in the reference population could affect the imputation accuracy. It should be noted that since the individuals were sorted based on their date of birth, increasing the number of individuals in the reference population also increased the average relationship between reference and target populations. We also showed more than 15,000 SNPs and 2,000 individuals in the reference were required to achieve more than 90 percent imputation accuracy for both individuals and SNPs when the pedigree information was not used and SNPs were selected randomly.

ACKNOWLEDGEMENTS

MHF, NKC and this study was supported by Meat and Livestock Australia project L.GEN.0174. The authors thank Angus Australia and Angus New Zealand for providing data for this study.

REFERENCES

Chadeau-Hyam M., Hoggart C., O'Reilly P., Whittaker J., De Iorio M. and Balding D. (2008) *BMC Bioinformatics* **9**: 364.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A. and Goddard M.E. (2012) *J. Dairy Sci*. **95**: 4114.

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

Connors N., Cook J., Girard C., Tier B., Gore K., Johnston D. and Ferdosi M. (2017) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **22:**

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

VanRaden P.M. (2008) *J Dairy Sci* **91:** 4414.