Proc. Assoc. Advmt. Anim. Breed. Genet. 20:554-557

UTILITY OF IMPUTED SNP GENOTYPES FOR GENOME-WIDE ASSOCIATION STUDIES IN DAIRY CATTLE

M.S. Khatkar, P.C. Thomson and H.W. Raadsma

ReproGen, Faculty of Veterinary Science, University of Sydney, Camden, NSW 2570

SUMMARY

Comparisons of genome-wide association studies (GWAS) based on imputed and actual genotypes were made using a dataset of 2,205 dairy bulls genotyped with a 50K SNP chip. The animals were divided into a reference (25 %) and a test panel (75 %). The genotypes of the test animals specific to two commercial lower density chips (*i.e.* 3K and 7K) were imputed up to the 50K using the IMPUTE2 software. The 'best guess' genotypes and allele dosages (estimated number of copies of an allele) were used as imputed genotypes. The association of SNP genotypes with phenotypes were conducted on five dairy traits (*viz*. milk yield, fat yield, protein yield, survival and daughter fertility) using true and imputed 50K genotypes of the test animals. The accuracy of imputation had a clear impact on the ability to detect the significant associations but varied between the 3K and the 7K, and among the five traits. The allele dosage model was superior to the best-guess model. Filtering the SNPs based on an indirect indicator of accuracy of imputation significantly improved the repeatability of GWAS results obtained from the imputed genotypes. Overall our results show that imputed genotypes can be used effectively to increase the power of GWAS.

INTRODUCTION

A number of SNP chips varying in SNP density and cost are available for genotyping cattle. For the dairy industry, an attractive strategy to increase genotypic information in a population whilst keeping cost of genotyping down is to genotype a large number of animals with a cheaper low-density SNP chip and impute up to high density genotypes using a limited number of reference animals genotyped with a high-density SNP chip (Khatkar *et al.* 2012). In addition to the primary utility of using imputed genotypes for genomic selection, such high-density imputed SNP genotypes on a large number of animals can boost the power of genome-wide association studies (GWAS) and fine-mapping of causal variants (Marchini and Howie 2010). GWAS rely on linkage disequilibrium (LD) between genotyped SNPs and causal mutations and hence benefit from the availability of very high-density SNP panels genotyped on large numbers of animals. In addition, genotype imputation is becoming a popular approach for combining multiple resource populations genotyped using different SNP panels, especially for meta-analysis (de Bakker *et al.* 2008; Jiao *et al.* 2011).

Imputation of genotypes is generally achieved with some uncertainty which may affect the ability to detect SNP associations. A number of studies have examined the accuracy and utility of imputed genotypes for GWAS in human (Marchini and Howie 2010). However, to our knowledge no study has been undertaken in livestock. The population structure, traits and density of the SNP panels in use in livestock are quite different from those in human. Such an analysis would provide useful information for conducting GWAS on imputed genotypes in cattle. Here we compared GWAS based on imputed and actual genotypes using a dataset of dairy cattle genotyped with a 50K SNP chip. We compared two types of imputed genotypes *viz.* 'best guess' and 'allele dosage', and investigated the effect of imputation accuracy on the repeatability of SNP association tests.

MATERIAL AND METHODS

Data. A total of 2,205 bulls genotyped with the Illumina BovineSNP50 chip were used in this

study (Khatkar *et al.* 2012). After filtering the SNP for low minor allele frequency (MAF>1%) and other QC measures, a total of 41,864 SNPs mapped on autosomes on UMD3.0 were used in this study.

Imputation. The 2,205 animals were divided into a reference (25 %) and a test panel (75 %). The 550 animals in the reference panel were selected randomly from the animals born before 2001 and all remaining animals were included in the test panel. For the reference panel, all the 50K genotypes were used. For the test panel a subset of the 50K SNP genotypes specific to two commercial lower density chips, *viz.* 3K (Wiggans *et al.* 2012) and 7K (Boichard *et al.* 2012), were used. Most of the SNPs on the 3K and the 7K chips are present on the 50K chip. The genotypes of the test animals were imputed up to the 50K using the genotypes of the reference animals. IMPUTE2 version 2.1.2 (Howie *et al.* 2009) was used for imputation. The 'best guess' genotype and the allele dosage were used as imputed genotypes. Allele dosage is the expected count of the B-allele.

Accuracy of imputation. Correlations between the actual and imputed genotypes were computed for each SNP by coding the AA, AB, BB genotypes as 0, 1, 2. In addition mean allelic error rates for the imputed genotypes were computed as the percentage of incorrectly predicted alleles *i.e.* mean allelic error rate (%) = number of incorrectly predicted alleles / total number of alleles imputed in the test set × 100.

SNP association. Association of SNP genotypes with five dairy traits (daughter trait deviations, DTD) were computed using the actual 50K and imputed 50K genotypes of the test animals. The five traits analysed were milk yield, fat yield, protein yield, survival and daughter fertility index which reflect a range of heritabilities. The regression of the traits on SNP genotypes were conducted by fitting the SNP allele count or allele dosage as a covariate and animal additive genetic effect as a random effect in a linear mixed model using ASReml (Gilmour, 2009). In addition each observation was weighted with the accuracy of DTD of each bull. The correlation of $-\log_{10}$(*p*-values) obtained by original 50K *vs.* imputed 50K was taken as the accuracy/ repeatability of GWAS on imputed genotypes for each trait.

RESULTS AND DISCUSSION

Overall agreement of SNP genotype association with milk volume as obtained using original and imputed genotypes *i.e.* best guess genotypes and allele dosage is presented in Figure 1. These results are based on imputed genotypes obtained by using the 3K SNP chip on the test animals. The repeatability of the *p*-values obtained using imputed allele dosage (0.92) was higher than the repeatability using best guess genotypes (0.89). Similar results were observed for other traits and when using the 7K SNP chip (results not shown). Higher repeatability using allele dosage could be expected as the probabilities of calling correct genotypes by imputation are included in the computation of allele dosage.

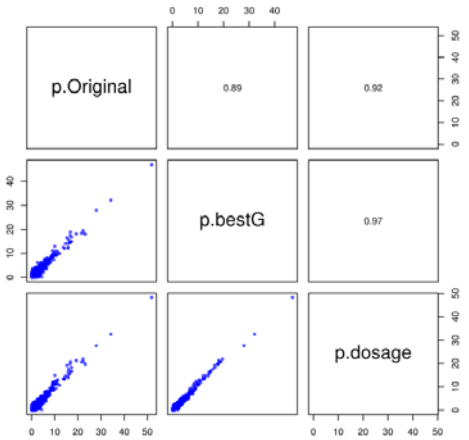

Figure 1. The repeatability of SNP associations with milk volume using imputed genotypes. The values in the upper triangle are Pearson correlation coefficients between -$\log_{10}$(*p*-values) using respective genotypes.

The results presented in Table 1 show further details on the repeatability of SNP association with phenotypes, where SNP genotypes were allele dosages from imputation. The correlation between -log10 of *p*-values varies from 0.84 to 0.92 across five traits. To examine the effect of accuracy of imputation on repeatability of association, the SNPs were classified according to their imputation accuracy. The SNPs with imputation accuracies less than 0.75 have low repeatability (Table 1). These results suggest that imputed genotypes of the SNPs with high error rate may not be useful for GWAS.

Table 1. The agreement of *p*-values for GWAS for five different traits obtained using actual genotypes and imputed genotypes (allele dosages)

Imputation accuracy (range)	n snp	MAER	Imputation accuracy	Repeatability of p-values				
				Milk volume	Fat	Protein	Direct survival	Cow fertility
ALL	39226	3.589	0.902	0.918	0.904	0.879	0.835	0.841
0.0-0.5	87	12.943	0.345	0.120	-0.04	0.104	0.227	0.366
0.5-0.75	1025	5.108	0.688	0.529	0.476	0.586	0.569	0.526
0.75-0.9	12484	4.786	0.857	0.860	0.821	0.808	0.757	0.779
0.9-0.95	19971	3.246	0.927	0.945	0.932	0.912	0.865	0.878
0.95-1.0	5659	1.738	0.963	0.947	0.951	0.946	0.934	0.927

Imputation accuracy is the correlation coefficient between imputed dosage and true genotypes; Repeatability of *p*-values = cor($-\log_{10}$(*p*-values- actual), $-\log_{10}$(*p*-values- imputed)); MAER = mean allelic error rate (%).

The accuracy of the imputation of untyped SNPs cannot be estimated in the absence of any true genotypes for comparison. However, it is possible to have some indication of quality of imputed genotypes. Browning and Browning (2009) suggested using the Pearson correlation between best guess and allele dosage as an indicator of accuracy of imputation. Figure 2 shows the relationship of this indicator with the accuracy of imputation. These results suggest that a large proportion of the SNPs with low accuracy of imputation can be filtered out by using the correlation between best guess and allele dosage as indirect measures. Such a filtering step can significantly improve the results of GWAS obtained from imputed genotypes.

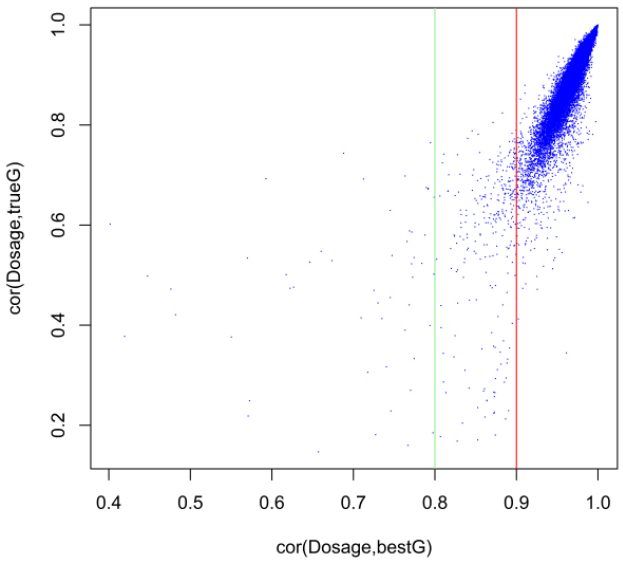

Figure 2. The relationship of correlation between allele dosage and best guess (x-axis) with the accuracy of imputation (y-axis; correlation between dosage and true genotypes).

The main motivation for undertaking a GWAS is usually to identify signals for causal variants or SNP in LD with such variants. Because of high LD between SNPs, especially when high-density SNP chips are used, true signals are generally represented by multiple SNPs in the region. The repeatability of individual SNPs from imputed genotypes is variable as discussed above, however, when the number of SNPs in a sliding window was used to detect the signal, the repeatability of signals using imputed GWAS was higher (results not shown).

Low MAF also affects the accuracy of imputation (Khatkar *et al.* 2012) and hence accuracy of association in GWAS. We excluded all SNPs with a MAF less than 1%. Excluding SNPs with very low MAF and filtering with the indicator of accuracy of imputation (Figure 2) can improve the GWAS results obtained from imputed genotypes.

We only tested the additive genetic effect of the SNP allele. It is possible to use the data on the cows to estimate the dominance effect by contrasting the mean of three genotypes. Such analysis will require using best guess imputed genotypes. With the availability of different SNP panels for bovine, it is becoming common place to genotype the same or different resource populations with different SNP chips. Imputation can help to combine such datasets. Recently we showed that the genotypes of animals can be imputed from 50K to 800K with a very small loss of accuracy of imputation (Khatkar *et al.* 2012). Such high-density imputed datasets will provide resources to conduct very powerful GWAS whilst maintaining the cost of genotyping at a low level.

ACKNOWLEDGEMENT

The data resource was generated by the CRC-IDP, the Department of Primary Industries, Victoria and the University of Sydney, with input from Drs Ben Hayes, Phil Bowman, Amanda Chamberlain, Matthew Hobbs and Mrs Gina Attard. The semen samples were provided by Genetics Australia and the phenotypic data by Australian Dairy Herd Improvement Scheme.

REFERENCES

Boichard D., Chung H., Dassonneville R., David X., Eggen A., Fritz S., Gietzen K.J., Hayes B.J., Lawley C.T., Sonstegard T.S. , et al.(2012) *PLoS ONE* **7**: e34130.

Browning B.L. and Browning S.R. (2009) *Am J Hum Genet* **84**: 210.

de Bakker P.I., Ferreira M.A., Jia X., Neale B.M., Raychaudhuri S. and Voight B.F. (2008) *Hum Mol Genet* **17**(R2): R122.

Gilmour A.R., Gogel B.J., Cullis B.R., Thompson R. (2009) ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK http://www.vsni.co.uk.

Howie B.N., Donnelly P. and Marchini J. (2009) *PLoS Genet* **5**: e1000529.

Jiao S., Hsu L., Hutter C.M. and Peters U. (2011) *Genet Epidemiol* **35**: 597.

Khatkar M.S., Moser G., Hayes B.J. and Raadsma H.W. (2012) *BMC Genomics* **13**: 538.

Marchini J. and Howie B. (2010) *Nat Rev Genet* **11**: 499.

Wiggans G.R., Cooper T.A., Vanraden P.M., Olson K.M. and Tooker M.E. (2012) *J Dairy Sci* **95**: 1552.