Proc. Assoc. Advmt. Anim. Breed. Genet. 20:550-553

ACCURACY OF IMPUTATION IN A POPULATION OF TROPICAL COMPOSITE CATTLE WITH PARTICULAR EMPHASIS ON THE USE OF ALLELIC R^2 AS A QUALITY CONTROL METRIC

M. Kelly, M.R.S. Fortes and S.S. Moore

The University of Queensland, Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, Brisbane, QLD 4062

SUMMARY

Imputation of genotypes from low-density single nucleotide polymorphism (SNP) panels to higher density panels is a common approach applied to increase the density of genotypes used in genomic selection and genome wide association studies (GWAS). Accuracy of imputation from Illumina BovineSNP50 to Illumina HD SNP panels was assessed within tropical composite beef cattle using 589 animals. The average imputation accuracy was high according to the percentage of concordant genotype calls (CONCORD) (96%) or the correlation between actual and imputed genotypes ($r^2_{(a,i)}$)(0.94). Considering imputed genotypes for a genome wide association study, we estimated that on average the power of GWAS to be approximately 12% less than genotyping all animals on Illumina HD. The accuracy of imputing individual SNPs was found to vary substantially, depending on multiple factors such as minor allele frequency and chromosome. There was a large number of SNPs for which the $r^2_{(a,i)}$ was less than 0.9. The allelic R^2 statistic reported by BEAGLE was able to identify a large number of such SNP. Placing a threshold on allelic R^2 statistic resulted in a marginal increase in average correlation between actual and imputed genotypes but a large decrease in the percentage of SNP with $r^2_{(a,i)}$ less than 0.81 (from 14% to 2.4%)

INTRODUCTION

Imputation of genotypes across different single nucleotide polymorphism (SNP) panels or from low density panels to high density panels is a routine way of increasing the number of markers for genomic selection (GS) and genome wide association studies in livestock. On average imputation accuracy is high and so genomic breeding values developed on imputed or actual genotypes are highly correlated (Brondum *et al.* 2012, Erbe *et al.* 2012). The impact of imputation accuracy on GWAS is less well understood. Additionally, the impact of using imputation on a diverse multi-breed reference population, such as the Tropical Composite beef cattle from northern Australia merits investigation. Breed diversity may have a negative impact on imputation accuracy and therefore it may affect both GWAS.

The aims of this study were: 1) to test the accuracy of imputation in a population of tropical composite beef cattle, 2) to test the effectiveness of using quality control statistics as a threshold for removing poorly imputed SNPs.

MATERIALS AND METHODS

Animals and genotypes. Genotype data from 589 Tropical Composite animals were used in this analysis. The Tropical Composite cattle consisted of both crossbred cattle and stabilised crosses from a range of founder breeds. Details on management and breeding of this cattle population developed by the Cooperative Research Centre for Beef Genetic Technologies are provided elsewhere (Barwick *et al.* 2009, Burns *et al.* 2013, Corbet *et al.* 2013). The Illumina HD bead chip was used to genotype the samples according to the manufacturer's protocols (Illumina Inc., San Diego, CA). Standard quality control: SNPs with call rate < 0.9 or

minor allele frequency < 0.01 were excluded. Missing genotypes were imputed using BEAGLE 3.2 (Browning and Browning, 2009). Quality control and imputation for missing genotypes resulted in 729,068 SNP with complete genotypes for 589 cattle.

Imputation from low density SNP panel. Imputation from the intersecting SNPs from Illumina BovineSNP50 to Illumina HD (729,068 SNP) was performed using the default settings in BEAGLE (Browning and Browning 2009). A 30 fold cross validation was used to ensure that the reference set of genotypes used to impute new genotypes was representative of the full reference population. The cross validation was performed in 3 steps as follows: 1) Groups of 20 animals were randomly allocated into 30 cross validation sets, 2) One set of 20 animals was imputed from BovineSNP50 to Illumina HD using the remaining groups HD SNP as reference genotypes, 3) this process was performed 30 times so each group had been used as a test set once.

Imputation accuracy and analysis. The accuracy of imputation calculated across animals within SNP was assessed two ways: 1) the concordance between actual and imputed genotype calls (CONCORD) and 2) the correlation between actual and imputed number of copies of the Allele coded B according to Illumina's A/B coding convention ($r^2_{(t,i)}$). The correlation was used as the primary statistic for assessing imputation as it is less influenced by minor allele frequency (Browning and Browning 2009). When imputing data generally we do not know the true accuracy of imputation for each SNP, BEAGLE provide a statistic called the allelic r^2 (R^2_{est}) which estimates the squared correlation between actual and imputed SNP. The effectiveness of this measure in identifying SNP with low CONCORD and ($r^2_{(a,i)}$) was assessed.

RESULTS AND DISCUSSION

On average imputation was good with a concordance rate of 0.96 and a ($r^2_{(a,i)}$) of 0.88 (Table 1). Thus the power of performing GWAS using imputed genotypes would be approximately 12% lower than using Illumina HD genotypes.

Table 1 Summary of concordance and correlation between actual and imputed genotypes with an increasingly stringent threshold applied using allelic r^2

Threshold on R^2_{est}	CONCORD	$r^2_{(t,i)}$	Markers excluded (%)
0	0.96	0.88	0.0
0.5	0.96	0.89	1.7
0.75	0.96	0.90	6.7
0.95	0.99	0.97	70.7

The measures of imputation accuracy in Table 1 are comparable with other studies performed in cattle with Erbe *et al.* (2012) finding concordance of actual and imputed genotypes of 0.97 in Holsteins and 0.96 in Jersey cattle. Present results were on the lower range of correlations between actual and imputed genotypes of 0.92-0.98, reported by Brondum *et al.* (2012). A slight reduction in imputation accuracy may be expected in the current study due to diverse genetic background of the cattle under investigation. Although the average imputation accuracy was quite high there was substantial variation in imputation accuracy. Imputation accuracy was affected by a number of factors including minor allele frequency and chromosomes, chromosome X in particular was imputed with lower accuracy.

As the threshold on R^2_{est} for excluding SNP became increasingly stringent the mean CONCORD and $r^2_{(a,i)}$ were high their means increased from 0.961 to 0.986 and 0.88 to 0.97 for CONCORD and $r^2_{(t,i)}$ respectively (Table 1). The editing of SNP based on R^2_{est} also decreased the number of SNP with low call rates, this is demonstrated visually in Figure 1 where fewer SNP with low $r^2_{(t,i)}$ appear successively from (a) through to (d). Additionally R^2_{est} was highly correlated with $r^2_{(a,i)}$(0.81).

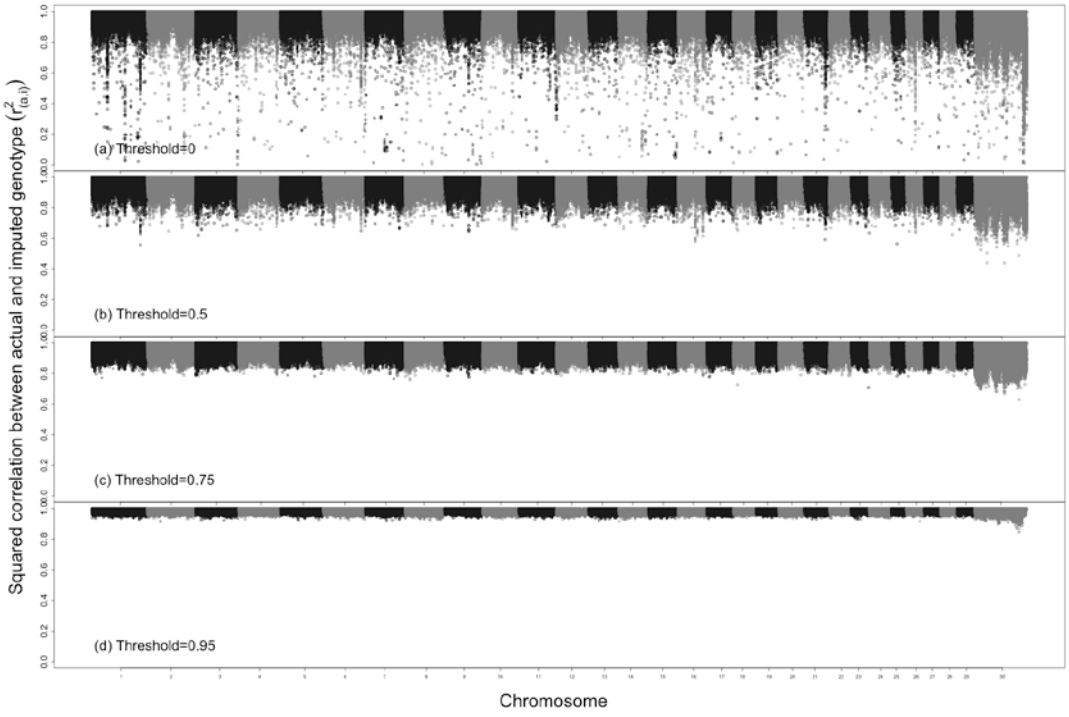

Figure 1. The correlations between imputed and actual genotypes with increasingly stringent thresholds applied using allelic r^2.

Figure 1 shows that many SNPs were imputed with low accuracy. The ability to identify such SNPs was examined by considering the ability of R^2_{est} to identify SNPs with CONCORD or $r_{(a,i)}$ lower than 0.9. False negatives were be defined as SNPs with correlation or concordance lower than 0.9 that were not excluded by quality control. Conversely, false positives would be SNPs with correlation or concordance greater than 0.9 that were excluded. As the R^2_{est} threshold for selecting SNPs becomes more stringent the number of false negatives decreases substantially (Table 2). There is a trade off as the number of false positives also increases, this is especially evident when the threshold is 0.9 or above. A reasonable compromise is to set the threshold to approximately 0.8 where false negatives (for the correlation) are reduced from 14.8% to 2.4% while false positives are 9.4%.

Table 2 Percentage of false negatives and false positives for concordance and correlation with an increasingly stringent threshold applied to BEAGLE r^2

	Percentage false negatives[*1]		Percentage false positives[*2]	
Threshold on allelic r^2	Concordance	Correlation	Concordance	Correlation
0	3.7	14.8	0.0	0.0
0.5	3.6	13.6	1.6	0.4
0.75	2.9	10.5	5.9	2.0
0.95	0.0	0.0	69.6	65.6

[*1] Percentage false negatives: percentage of SNPs with correlation or concordance lower than 0.9 that were not excluded; [*2]Percentage of false positives: percentage of SNPs with correlation or concordance greater than 0.9 that were excluded.

The current study focused on a small part of genotype quality control for use of imputed genotypes in GWAS studies. Attention must be played to quality control at all stages of the analysis. The detection of imputation accuracy per individual animal would also be an important step to improve the overall quality control. It was found that the genotype probability of each genotype call averaged over each animal was not related to overall imputation accuracy (data not shown).. In summary, special consideration of individual SNP imputation accuracy could avoid detection of false QTL, when performing genome wide associations with imputed SNP data. It is possible to use R^2_{est} as a quality control statistic to reduce imputation accuracy issues.

ACKNOWLEDGEMENTS

Northern Pastoral Group, Department of Employment, Economic Development and Innovation (DEEDI) and CSIRO Animal Food and Health Science. Animal Genetics and Breeding Unit in particular Bruce Tier. Meat and Livestock Australia, Australian Centre for International Agricultural Research, Cooperative Research Centre for Beef Genetic Technologies, Northern Pastoral Group, CSIRO – Animal Food and Health Science, DEEDI and The University of Queensland gave financial support.

REFERENCES

Barwick S. A., Johnston D. J., Burrow H. M., Holroyd R. G., Fordyce G., Wolcott M. L., Sim W. D. and Sullivan M. T. (2009) *Animal Production Science* **49**: 367.

Brondum R. F., Ma P., Lund M. S. and Su G. (2012) *J Dairy Sci* **95**: 6795.

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

Burns B. M., Corbet N. J., Corbet D. H., Crisp J. M., Venus B. K., Johnston D. J., Li Y., McGowan M. R. and Holroyd R. G. (2013) *Animal Production Science* **53**: 87.

Corbet N. J., Burns B. M., Johnston D. J., Wolcott M. L., Corbet D. H., Venus B. K., Li Y., McGowan M. R. and Holroyd R. G. (2013) *Animal Production Science* **53**: 101.

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.