ACCURACY OF SNP IMPUTATION USING A MULTI-TIERED GENOTYPING APPROACH IN DAIRY CATTLE

M.S. Khatkar[1,3], G. Moser[1,3], B. Hayes[2,3] and H.W. Raadsma[1,3]

[1]ReproGen Animal Bio-science group, University of Sydney, Camden, NSW 2570
[2]Biosciences Research Division, Department of Primary Industries Victoria 3083
[3]Dairy Futures Cooperative Research Centre (CRC), Australia

SUMMARY

Accuracies of different imputation strategies to impute genotype data for untyped or masked SNPs were explored using data on 2,727 animals genotyped with the Illumina BovineSNP50 BeadChip. Various 2-tier and 3-tier imputation scenarios with reference panels of varying sizes and marker densities were generated, and compared by masking the known genotypes in the test panel. The accuracy of imputation increased as the number of animals in the reference panel increased and the SNP density of the test panel increased. For animals genotyped with a low density panel, there was a gain in accuracy of imputation from 0.5 % to 7 % in a 3-tiered approach using a combination of high and medium and low density reference panels, over a 2-tiered approach using only low density and high density panels. The implications for use of ultra-high density SNP panels and whole genome sequence content are discussed.

INTRODUCTION

Genotyping with high density SNP panels (chips) is important for accurate prediction of phenotypes and Direct Genomic Values (DGV). Very high density SNP panels and whole genome sequencing is becoming readily available in a number of species. A number of SNP chips have been developed in cattle which includes 15k (Khatkar *et al.* 2007), 25k (Raadsma *et al.* 2009), 50k (Matukumalli *et al.* 2009) and more recently 650k (http://www.affymetrix.com) and 800k (http://www.illumina.com). These SNP chips have now been widely used for genotyping a number of bovine populations. As new chips are developed, re-genotyping previously genotyped samples or new samples for whole genome sequencing or very high density SNPs is expensive. A more cost effective approach, would be to genotype a small proportion of the population using a high-density SNP panel and then employ genotype imputation methods for predicting high-density genotypes for the rest of the population genotyped with a lower density and lower cost SNP panel.

Genotypic imputation is defined as the prediction of genotypes at the SNP locations for which assays are not directly available, in a sample of individuals. There are many scenarios where imputation can be used. Imputation in this study refers to the situation in which one or more a reference panel of animals is genotyped with a set of higher density SNP chips and is used to predict the genotypes of test samples that have been genotyped with a subset of these SNPs. The *in silico* genotypes obtained by imputation can then be used in genome wide association and genomic selection analyses (Browning and Browning 2007; Goddard and Hayes, 2009). Such strategies are likely to result in more accurate predictions of DGV, and improve the ability to resolve or fine-map QTL or QTN, and facilitates meta-analysis across larger data sets with heterogeneous SNP information

A number of imputation programs (fastPHASE (Scheet and Stephens 2006), MACH (Willer *et al.* 2008), IMPUTE (Howie et al. 2009), Beagle (Browning and Browning 2007) allow imputation of genotypes. Accuracy of imputing of sporadic missing genotypes that occur when calling genotypes from genotyping chips, is often very high. The present study aimed to infer genotypes at untyped markers (systematic missing data) using various reference panels. IMPUTE accommodates the use of different reference panels in a tiered or staged fashion. IMPUTE has also been demonstrated to achieve a high accuracy of imputation (e.g.Weigel et al., 2010); hence we chose this method to examine the performance of imputation under various scenarios by varying the size and SNP density of the reference and test panels.

MATERIAL AND METHODS

Genotype data: The genotypic data on 2,727 animals (2,205 bulls and 522 cows) genotyped (Moser *et al*. 2010) with the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, USA)) were used for this study. After quality control, a total of 1324 SNPs on chromosome 20 were used for the present analyses.

Imputation methods: We utilized IMPUTE program which is based on an extension of the hidden Markov models, and uses a fine-scale recombination map across the genome (Howie *et al*. 2009). IMPUTE provides the probability of different possible genotypes at each missing genotype. We used the best-guess genotype as predicted genotype for comparing the accuracies under different scenarios. The accuracy of imputation was computed as the percentage of correctly predicted genotypes, and error rate as the percentage of incorrectly predicted genotypes.

Imputation Scenarios: Two imputation strategies (2-tier and 3-tier) were compared. In the 2-tier a single reference panel with higher density SNPs was used to impute the genotypes in the test panel genotyped with a lower density SNP panel. In the 3-tier approach two reference panels were used; a top or main reference panel genotyped with high-density SNPs and a middle panel with medium-density SNPs and a test panel genotyped with a low- density SNP panel. Three sizes of top reference panels were generated by randomly selecting 27, 136 or 270 bulls representing 1, 5 and 10 % of total samples. Two sizes of middle panels consisting of 10 % and 50 % of the total samples were tested. A set of evenly spaced 611 SNPs, equivalent to 20k genome wide SNPs, was used for middle reference panel. Two densities of SNPs for test panels representing a genome wide 3k and 5k were explored. These SNP densities were generated by iterative thinning the SNPs based on spacing and retaining SNPs with higher minor allelic frequency (MAF). The combination of the size of panels and density of SNPs under the different scenarios are presented in Table 1 and Table 2.

RESULTS AND DISCUSSION

Accuracies of imputation for different imputation scenarios using the 2-tier approach (scenario 1 to 8) are given in Table 1 and for the 3-tier strategy (scenario 9 to 15) in Table 2. For both strategies, the accuracy of imputation increased with the size of the reference panel. The accuracy of imputation increased from 82.1 % (scenario 9) to 92.7 % (scenario 11) when the reference sample was increased from 27 to 270 bulls for the 2-tier approach (Table 2).

Accuracy of imputation was higher under the 3-tier strategy in all the scenarios which were directly comparable to the same scenarios under the 2-tier approach. The accuracy of imputation in scenario 1 (2-tier) using a single reference panel of 27 bulls was 82.1 % (Table 1), and the accuracy increased by more than 7 % over scenario 1 when an additional panel of medium density SNPs was included in a 3-tier framework (scenario 8 & 9, Table 2). The additional gain in accuracy was smaller when the number of bulls in the top panel was increased. For example, the gain in accuracy was only 0.6 % under scenario 11 as compared to Scenario 3 where the top panel had 272 bulls (10 % of the samples). Similar observations were made when the test panel had 144 SNP (equivalent of 5k genome wide density). However, the gain in the accuracies in 3-tier over 2-tier were slightly less for the 5k test panel compared to the 3k test panel. For example there was a

6.2 % increase in accuracy of scenario 13 (3-tier) over scenario 4 (2-tier). This shows, that as the SNP density in test panel increases, the additional gain of using 3-tier approach becomes smaller. Highest accuracy of imputation (97.4 %) was obtained under scenario 7 with largest panel of reference bulls and a 20k medium density test panel.

Table 1. Accuracy of imputation under different scenarios using 2-tier approach

	Reference Panel		Test Panel			
Scenario	n animals	n snp (50k)	n animals	n snp	chip	Accuracy
1	27	1324	2700	85	3k	82.1
2	136	1324	2591	85	3k	90.6
3	272	1324	2455	85	3k	92.7
4	27	1324	2700	144	5k	84.8
5	136	1324	2591	144	5k	92.7
6	272	1324	2455	144	5k	94.7
7	272	1324	2455	611	20k	97.4

Table 2. Accuracy of imputation under different scenarios using 3-tier approach

	Top Reference Panel		Middle Reference Panel		Test Panel			
Scenario	n animals	n snp (50k)	n animals	n snp (20k)	n animals	n snp	chip	Accuracy
8	27	1324	270	611	2430	85	3k	89.1
9	27	1324	1347	611	1353	85	3k	89.3
10	136	1324	1279	611	1312	85	3k	92.3
11	272	1324	1186	611	1269	85	3k	93.3
12	27	1324	270	611	2430	144	5k	90.9
13	27	1324	1347	611	1353	144	5k	91.0
14	136	1324	1279	611	1312	144	5k	94.1
15	272	1324	1186	611	1269	144	5k	95.2

We also investigated the effect of minor allelic frequencies (MAF) of the masked SNPs on the accuracy of imputation. The error rate is higher and more variable when the MAF of SNP increases above 0.1 (Figure 1), which suggests that genotypes of common SNP are more difficult to impute. In general there is higher probability of sampling correct genotype for a SNP with lower MAF from the distribution of three genotypes. Impute uses information from adjacent SNPs to impute correct haplotypes. Hence, accurate imputation of common SNP may require higher density SNP panels in the test samples. There was no pattern of relationship of the error rate with the HWE test of the SNP (data not shown).

In this study we demonstrated that additional gains in accuracy of genotype imputation can be achieved by employing an additional reference panel of medium SNP density in the imputation process. This approach is in particular suited for situations where a small fraction of the population is genotyped for a high-cost ultra-high density assay or whole genome sequencing data is available, and a larger panel of samples is genotyped with medium-density SNP chip as now becoming available in cattle. Then very large numbers of routine field samples genotyped with a low-cost lower-density panel can be imputed for whole genome sequence using the reference panels in tiered fashion. These *in-silico* genotypes would contribute towards increased accuracy of genomic selection and increased genetic gains with the use of DGV.

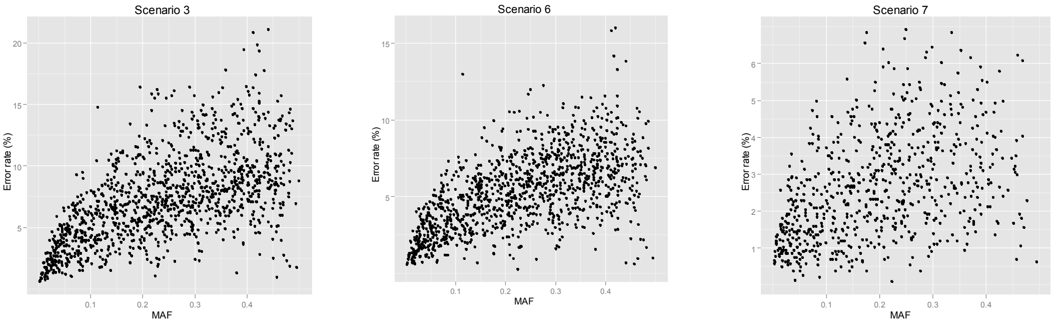

Figure 1. Comparison of imputation error rate versus MAF of SNP for imputation scenario 3, 6 and 7.

CONCLUSIONS

In this study we present the utility of IMPUTE as a genotype imputation method with varying sizes of reference panels and different SNP density. We showed that there is a gain in accuracy of imputation by including an intermediate reference panel in 3-tier (two reference panels) as compared to using 2-tier (single reference panel) especially when the reference panel is small. The accuracy of imputation is affected by the size of the reference panel, the density of SNP in the test panel and also by MAF of the imputed SNP.

ACKNOWLEDGEMENT

The research was funded by the Dairy Futures cooperative Research Centre. The data resource was generated by the CRC-IDP and the Department of Primary Industries, Victoria, with input from Drs, Phil Bowman, Amanda Chamberlain, Matthew Hobbs and Mrs Gina Attard. The semen samples were provided by Genetics Australia.

REFERENCES

Browning S.R. and Browning B.L. (2007). *Am J Hum Genet* **81**:1084.

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

Khatkar M.S., Zenger K.R., Hobbs M., Hawken R.J., Cavanagh J.A., Barris W., McClintock A.E., McClintock S., Thomson P.C., Tier B., Nicholas F.W. and Raadsma H.W. (2007) *Genetics* **176**:763.

Goddard M.E. and Hayes B.J. (2009) *Nat. Rev. Genet.***10**:381-91.

Matukumalli L.K., Lawley C.T., Schnabel R.D., Taylor J.F., Allan M.F., Heaton M.P., O'Connell J., Moore S.S., Smith T.P., Sonstegard T.S. and Van Tassell C.P. (2009) *PLoS ON*E **4**:e5350.

Moser G., Khatkar M., Hayes B. and Raadsma H. (2010). *Genet. Sel. Evol.* **42**:37.

Raadsma H.W., Khatkar M.S., Moser G., Hobbs M., Crump R.E., Cavanagh J.A. and Tier B. (2009) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18**:151.

Scheet P. and Stephens M. (2006). *Am J Hum Genet* **78**:629.