Proc. Assoc. Advmt. Anim. Breed. Genet. 19:379-382

IMPUTATION OF SINGLE NUCLEOTIDE POLYMORPHISM GENOTYPES IN A CROSSBRED DAIRY CATTLE POPULATION USING A REFERENCE PANEL

D.L. Johnson, R.J. Spelman, M.K. Hayr and M.D. Keehan

LIC, Private Bag 3016, Hamilton 3240, New Zealand

SUMMARY

Greater accuracy in the prediction of genomic breeding values may be achieved by the use of a high-density (HD) marker panel in order to increase the level of linkage disequilibrium between markers and quantitative trait loci (QTL). An objective was to evaluate, using a reference HD panel containing 700K markers, the accuracy of imputation of SNP markers in the HD panel that are not included in a panel of lower density. Results using a population-based algorithm suggest close to 99% accuracy for genotype imputation from a medium-density panel (50K) to a high-density panel (700K) and 96% accuracy for imputation from low-density (3K) to medium-density (50K).

INTRODUCTION

The availability of genome-wide dense marker maps has revolutionised dairy cattle breeding programs. Genomic breeding values are now being used in the dairy industry for bull selection. The basic principle of genomic selection (Meuwissen *et al.* 2001) is that QTL are in linkage disequilibrium (LD) with flanking markers and therefore the markers would be expected to explain a high proportion of the genetic variance if marker density is sufficiently high. Genomic selection implicitly uses both linkage analysis (LA) information, genetic relationships captured by markers, as well as LD which relates to information derived from chromosomal segments inherited from founder animals (Habier *et al.* 2007). Luan *et al.* (2010) showed that, based on the 50K panel, the contribution from LD, as opposed to LA, may be relatively small. Consequently this may be a barrier to the capture of Mendelian sampling variance for young bulls and also limit the ability to use markers across breeds where family relationships no longer hold.

Two factors influencing the accuracy of genomic predictions relate to the size of the reference or training population and to the density of the genetic markers. The latter will affect the level of LD between markers and QTL. The high-density SNP panel provides an option for increasing the marker density but the cost of this marker panel is currently too high to justify general use in dairy cattle breeding. One option is to use the HD panel on a reference group of individuals and to infer the missing genotypes for those individuals genotyped on the 50K panel (Goddard and Hayes 2009).

A low-density genotyping platform may be a low-cost option for use on commercial dairy farms for routine activities such as selection of replacement heifers. Imputation of remaining SNP up to the 50K level using genotypes of key ancestors may provide a low-density option that is applicable across traits and breeds (Habier *et al.* 2009). The objective of this study is to present results on the success of the genotype imputation between different densities of SNP panels.

MATERIALS AND METHODS

High density genotypes. A total of 2781 animals were genotyped with the Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA). This set included 147 bulls, 145 of which had previously been genotyped with the Illumina BovineSNP50 panel. Breed composition was 1261 Holstein-Friesian (HF), 1101 Jersey (JE), 374 Friesian-Jersey crossbreds (FJ) and 45 animals classed as other breeds and crosses. The number of markers retained in the HD set after quality control was 711,955 and 38,296 of these were present in the 50K subset. The animals were split into 2 groups at random. One group acted as the reference set which contained all markers and the other a test group where the markers not in the 50K subset were masked. The Beagle 3.2 software (Browning and Browning 2009) was used to impute the masked genotypes. Beagle uses an approach based on hidden Markov models to simultaneously phase and sort haplotypes into clusters. The population-based algorithm was used in the sense that individuals were assumed to be unrelated. The resulting genotype imputations from Beagle were then checked for pedigree consistency and, if there was a parent-progeny conflict, an imputed genotype was changed to the next most probable genotype based on the posterior genotype probabilities. The genotype imputations were then compared with the true genotypes to assess the imputation success rate in terms of both genotypes and alleles. The allelic R^2 measure of imputation accuracy, the squared correlation between the allele dosage (number of minor alleles) with the highest posterior probability and the true dosage (Browning and Browning, 2009), was used to eliminate poorly imputed markers prior to using the full HD set as the reference set in downstream analyses.

An initial group of 7256 animals that had previously been genotyped with the 50K panel were then imputed to HD level using the 2781 HD animals as the reference set. There were 145 animals in common between the 2 sets and these animals were retained in the 50K group to provide an additional check on imputation accuracy.

Imputation from 3K panel to 50K panel. A specialised low-density (3K) platform developed by Illumina (San Diego, CA) in cooperation with the Bovine Functional Genomics Laboratory (Beltsville, MD) was considered as the low-density option. This panel comprised 2977 markers. A total of 4356 bulls were genotyped with the 50K panel. For the 3 youngest cohorts of bulls (n=896) the genotypes were masked except for those markers in the 3K subset. Beagle 3.2 software was used as above for population-based imputation of the masked genotypes. In addition, sires with at least 10 progeny in the reference set were haplotyped using the rule-based method of Druet *et al.* (2008). The 146 derived haplotypes were then input to Beagle as phased genotypes in an attempt to increase imputation accuracy through the use of both linkage and linkage disequilibrium information. The BLUP estimation method (Meuwissen *et al.* 2001) was used to compare the correlations between predicted genomic breeding values and phenotype for the young bull test set. The test correlation was calculated for 3 scenarios: (i) train 50K, test 50K; (ii) train 50K, test 50K imputed from 3K; (iii) train 3K, test 3K. The phenotype was protein EBV for Holstein-Friesian bulls.

RESULTS AND DISCUSSION

High-density imputation. The average imputation success rate when masking a random half of the HD set was 98.96% for genotypes, ranging from 98.40% to 99.28% across chromosomes. Many of the errors still have one allele correct and the average allele imputation success rate was 99.47%. The frequency distribution of the proportion of masked genotypes that were imputed correctly on chromosome 1 is shown in Figure 1, the average and median genotype success rates were 99.15% and 99.48% respectively for this chromosome. The distribution for the Jersey breed appears to have a higher mode compared with other breeds but this may be due to a higher percentage of monomorphic loci for the Jersey breed as the success rate was based on markers with non-zero minor allele frequency (MAF) across breed. The allelic R^2 measure of accuracy as a function of MAF, when grouped into bins of size 0.01, is shown in Figure 2. The median allelic R^2 was greater than 0.97 for most MAF bins. The R^2 measure of imputation accuracy tends to increase with MAF.

On the basis of the allelic R^2 measure, 19,357 markers with $R^2 < 0.9$ were eliminated prior to using the full HD set of animals as the reference set. For the 145 bulls common to both panels, the average genotype imputation success rate was >99.9% and not much lower than the degree of concordance between markers common to the 50K and HD panels.

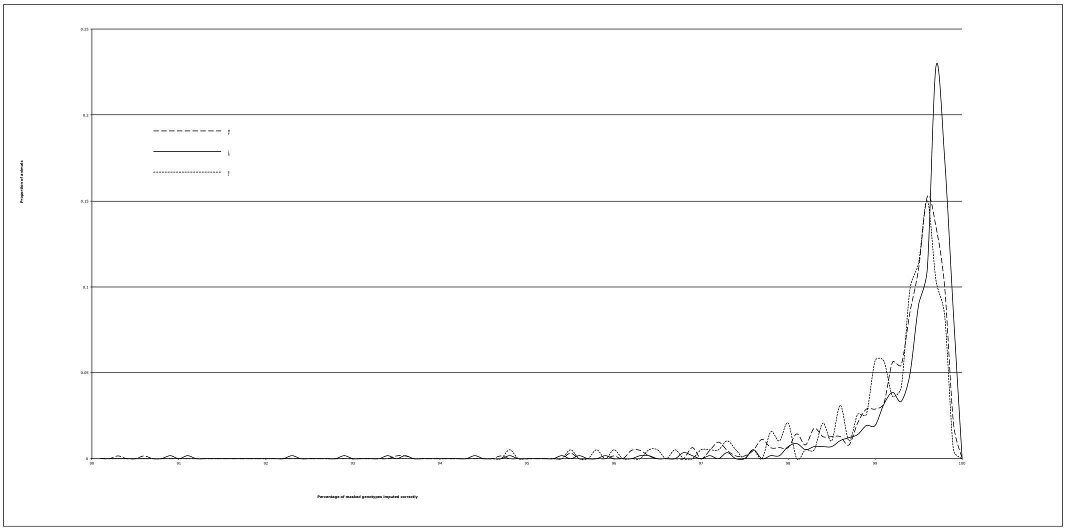

Figure 1. Frequency distribution of proportion of masked genotypes that were imputed correctly for high-density imputation on chromosome 1.

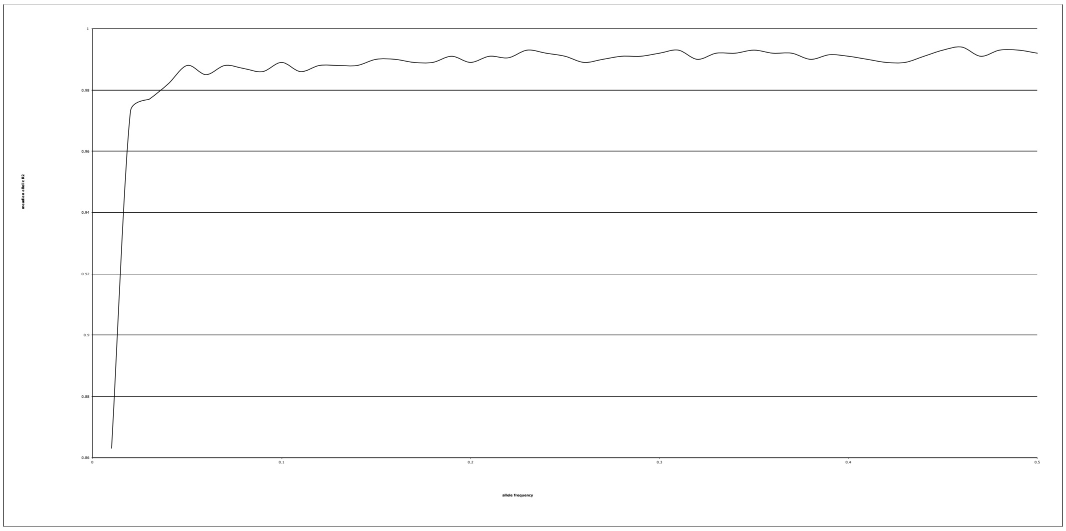

Figure 2. Median allelic R^2 and minor allele frequency for high-density imputation on chromosome 1.

Medium-density imputation. The average genotype imputation success rate was 96.93% for the young bulls when imputing from 3K to 50K. The variation across chromosomes was approximately ±1%. No improvement in accuracy was obtained by providing haplotype information on some proven sires suggesting that the population-based method used in Beagle was able to capture most of the relevant information. The test correlations for the BLUP analysis are shown in Table 1 indicating a loss of about 1% when using the imputed marker set compared to the full 50K set.

Table 1. Correlation between predicted genomic BV and protein phenotype for Holstein-Friesian bulls for different SNP marker panels.

Train SNP panel	Test SNP panel	Test correlation
50K	50K	0.566
50K	50K imputed from 3K	0.559
3K	3K	0.469

CONCLUSIONS

Incorporation of SNP panels of different densities into genomic evaluations combined with the utilisation of imputation techniques could greatly enhance the efficiency of breed improvement programs. Imputation from the 50K panel to the HD panel can be achieved with a high degree of accuracy with an average genotype success rate close to 99% . For the 3K imputation to 50K density the corresponding figure was close to 96%. In the latter case, the loss in accuracy of genomic breeding values due to using imputed markers compared with true values appears small. The success of the HD imputation will ultimately lie in the ability of the HD panel to improve the accuracy of prediction of genomic breeding values above the levels currently being achieved by the 50K panel.

REFERENCES

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

Druet T., Fritz M., Boussaha M., Ben-Jemaa S., Guillaume F., Derbala D., Zelenika D., Lechner D., Charon C., Boichard D., Gut I., Eggen A. and Gautier M. (2008) *Genetics* **178**: 2227.

Goddard M.E. and Hayes B.J. (2009) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18**: 26.

Habier D., Fernando R.L. and Dekkers J.C.M. (2007) *Genetics* **177**: 2389.

Habier D., Fernando R.L. and Dekkers J.C.M. (2009) *Genetics* **182**: 343.

Luan T., Woolliams J.A. and Meuwissen T.H.E. (2010) *Proc. 9th World Congr. Genet. Appl. Livest. Prod.*, Leipzig, Germany.

Meuwissen T.H.E., Hayes B.J. and Goddard M.E. (2001) *Genetics* **157**: 1819.