Prediction Of Missing Markers With Low Density Marker Panels In Dairy Cattle

Z. Zhang[*], M. Georges[*] and T. Druet[*]

Introduction

In dairy cattle, low density SNP panels have been proposed to genotype larger fraction of the population at lower costs (e.g. Habier et al., 2009). Animal breeders are now considering this technology and apply it to genomic selection. Basically, two options can be applied to predict breeding values for animals genotyped on low density chips: estimate the effects of the markers (or a combination of these markers) on the low density panel or predict the missing markers from a larger panel (imputation) and then apply genomic selection with the "complete" marker set.
In this study, we estimate the efficiency of the method described in Druet and Georges (2010) to estimate accuracy of marker imputation with low density marker panels.

Material and methods

Data. A set of 4738 animals genotyped for 45,836 SNP markers on the CRV chip, a custom-made 60K Illumina panel described in Charlier et al. (2008), was used in this study. For testing imputation efficiency, the animals were assigned to two groups: reference individuals were genotyped on all the markers while the target animals were genotyped on lower density SNP panels. All the 516 individuals with at least one genotyped offspring were considered as reference individuals. 484 additional animals were selected randomly as reference individuals. Finally, the remaining animals were assigned to the group of target animals.

Creation of low density marker panels. To mimic low density genotyping, different low density panels were defined (see below) and genotypes of target animals were erased for the unselected markers. Five different number of markers per chip were selected: 384, 768, 1536 (corresponding to minimal, intermediate and maximal number of markers with the Illumina Golden Gate technology - Illumina, San Diego, CA, USA), 3000 (corresponding to minimal number of markers with the Illumina Bead Chip - Illumina, San Diego, CA, USA) and 6000 markers. The markers were selected to obtain a compromise between uniform marker density and high minor allelic frequency (MAF) with the following method. Number of markers per chromosome was obtained by multiplying the desired marker density (total number of markers divided by the size of the genome) by the size of the chromosome. Each chromosome was divided in equal segments based on the desired number of markers. Then, the marker with the highest MAF was selected in the first segment. For subsequent segments,

[*] *Unit of Animal Genomics, GIGA-R B34, 1 avenue de l'Hôpital, B-4000 Liège, Belgium*

the marker with the highest score combining MAF and distance with the marker retained in the previous segment was selected:

$$score(i)=MAF(i)*[ssize - |ssize-dist(i)|]$$

where i is the indice of the tested marker, ssize is the size of each segment and dist(i) is the distance between the tested marker and the selected marker in the previous segment (Matukumalli et al., 2009).

Marker imputation method. Markers were imputed using Beagle (Browning and Browning, 2007) and DAGPHASE from the PHASEBOOK package (Druet and Georges, 2010). The method relies on linkage and linkage disequilibrium information. When a parent is genotyped, the corresponding offspring chromosome is modeled as a combination of the two parental haplotypes through linkage. When the parent is not genotyped, the corresponding chromosome is modeled with a directed acyclic graph (DAG) describing haplotypes from the population (Browning and Browning, 2007). This DAG models LD due to old ancestors (short range association) but also long range association due to more recent ancestors.
All genotypes from reference individuals were conserved while for target individuals only markers selected on the low density chips were used (the remaining markers were erased). Imputation efficiency was estimated by comparing imputed markers and real markers.

Results and discussion

Imputation efficiency by marker density. Figure 1 shows the relationship between mean imputation error rates (for all target animals) and number of selected markers on the low density SNP panel.

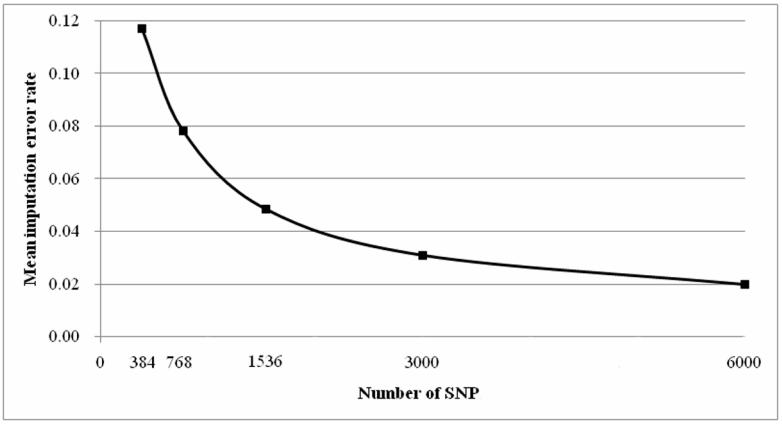

Figure 1: genome wide imputation error rates in function of number of SNP

The benefit of adding markers was decreasing at higher marker density. With less than 1000 SNP covering the genome, imputation error rates were above 5% (11.7% with 384 markers and 7.8% with 768 markers). From 1536 markers to 3000 markers, the error rate was still

reduced from 4.8% to 3.1% whereas from 3000 markers to 6000 markers, the mean error rate decreased only by 1.1%, resulting in 2.0% errors. In a study with higher density panels, it was shown that these figures can be improved to a certain extend by increasing the number of reference individuals (Druet *et al.* (2010)).The impact of these errors rate on accuracy of breeding values obtained by genomic selection should be estimated to conclude which low marker density panel would results in acceptable results for genomic selection. The optimal marker density must take into account both genotyping costs and accuracy of genomic selection.

Impact of relationship between target and reference individuals. The method directly models linkage between genotyped offsprings and parents. In addition, the DAG represents haplotypes of the reference individuals. Therefore, genotypes of target animals with genotyped parents or with more ancestors included in the reference individuals group should be predicted more accurately. To test this we computed a score representing the expected proportion of the genome of an animal which was inherited from a reference individual. It was equal to the mean score of parents (when a parent is genotyped, its score was replaced by one). The computed scores ranged from 0.0 to 1.0 with the median equal to 0.9375. Table 1 presents mean imputation error rate for animals with different proportion of ancestors genotyped (in the reference group). For all marker densities, the mean imputation error rate decreased asymptotically when the score increased. At lower marker densities (384 or 768 markers per chip), no differences were observed for scores between 0.85 and 0.99 whereas for higher marker densities, imputation rates where still decreasing for higher scores. This indicates that the benefit of having more ancestors genotyped is limited by the number of markers. Indeed, with fewer markers it is more difficult to identify the correct path of a haplotype in the DAG.

Table 1: imputation efficiency by score representing the expected proportion of the genome inherited from a reference individual

Score	Number of individuals	384	768	1536	3000	6000
<0.50	27	0.203	0.151	0.104	0.068	0.044
[0.50-0.75)	178	0.155	0.114	0.077	0.052	0.035
[0.75-0.80)	218	0.133	0.092	0.061	0.041	0.028
[0.80-0.85)	162	0.130	0.088	0.055	0.036	0.023
[0.85-0.90)	536	0.128	0.085	0.055	0.036	0.023
[0.90-0.95)	832	0.127	0.085	0.052	0.033	0.021
[0.95-1.00)	1279	0.128	0.085	0.051	0.031	0.019
1	506	0.032	0.017	0.009	0.005	0.003

For animals with both parents in the reference group (score = 1), the mean imputation rate was clearly lower than for other animals, ranging from 0.032 to 0.003 with 384 and 6000 SNP, respectively. This is explained by the fact that for these animals, both haplotypes are modeled through linkage and genotypes from parents are directly transmitted with high accuracy whereas for other animals, the maternal haplotype (and sometimes the paternal haplotype too) is modeled through the DAG. The imputation error rate per animal is the

mean from imputation error rates from paternal and maternal haplotype. Since sires are most often genotyped, we can assume that imputation efficiency for paternal haplotypes is approximately equal to the imputation efficiency observed for animals with both parents in the reference group (because all these haplotypes are described through linkage). Therefore, the imputation error rate for maternal haplotypes is probably higher (approximately equal to twice the error rate per animal minus the error rate observed when the score equals 1).

Conclusion

The results of this study show that with approximately 1 SNP / Mb (~3000 markers) missing genotypes can be predicted with ~3% errors. These figures can still be reduced with more markers. However the benefit of adding more markers will be limited at higher marker densities. It was also observed, that the imputation efficiency was larger when more ancestors where genotyped on the high density panel, particularly for animals with both parent genotyped at higher density. For these animals and with 3000 SNP or more per panel, the imputation rate was below 0.5%. The impact of these imputation rates on accuracy of genomic selection must be estimated to conclude which low marker density panel would be optimal. We tested also the efficiency of a method based only on linkage (Druet and Farnir, 2010). This last method performed better for animals with many reference individuals as ancestor, particularly at the lowest marker densities.

Acknowledgments

Tom Druet is Research Associate from the Fonds de la Recherche Scientifique – FNRS. The authors thank CRV for access to the data. This work was funded by grants of the Service Public de Wallonie and from the Communauté Française de Belgique (Biomod ARC). The authors acknowledge University of Liège (SEGI and GIGA bioinformatics platform) for the use of NIC3 and GIGA-grid supercomputers.

References

Browning, S., Browning, B., (2007). *Am J Hum Genet*., 81:1084-1097.

Charlier, C., Coppieters, W. and Georges, M., (2008). *Nature Genetics*., 40:449-454.

Druet,T., Georges, M., (2010). *Genetics*., (in press).

Druet, T., Farnir, F., (2010). In preparation.

Druet, T., Schrooten, C. and De Roos, A.P.W, (2010). *Proc 9th WCGALP*.

Habier, D., Fernando, R . and Dekkers, J., (2009). *Genetics*., 182: 343–353.

Matukumalli, K., Lawley, T. and Van Tassell, P., (2009). *Plos One*., 4(4): e5350.