Accuracy Of Direct Genomic Values In Holstein Cows Using Subsets Of SNP Markers

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

Introduction

In genomic selection (GS), selection decisions are made on genomic breeding values predicted from high-density Single Nucleotide Polymorphic (SNP) markers. Schaeffer (2006) demonstrated that significant additional gains could be made using GS for cows to breed sires and cows to breed cows. However at the current price, high density SNP genotyping arrays may be limited to applications involving elite sires and dams. An alternative is to use a lower density array for GS in the cow selection pathways. As shown for a single trait by Weigel et al. (2009), a low-density assay comprising selected SNP can deliver a substantial portion of the gain possibly for fraction of the price. However, such a low-density array may still be of limited use if multiple traits require so many SNP that the cost of genotyping them is similar to the cost of a high-density chip. The objective of this study was to evaluate the potential of low-density SNP genotyping assays to predict direct genetic value (DGV) in cows for four traits.

Material and methods

Phenotype and genotype data. SNP genotypes were derived from the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, USA). After quality control and omitting SNP located on the sex chromosomes, a total of 42,576 markers remained for the analysis. The phenotypes used were deregressed Australian Breeding Values (ABV) for protein percentage, Australian Selection index (ASI, a profit index) and survival index and daughter trait deviations for overall type, taken from the August 2009 Australian Dairy Herd Improvement Scheme (ADHIS; http://www.adhis.com.au/) evaluation. Accuracies of ABVs (square root of the published reliability) ranged from 0.82 to 0.95 for bulls and from 0.57 to 0.71 for cows and accuracies were higher for traits with higher heritability. The analyses included 2,142 Australian Holstein-Frisian bulls and 522 cows, except for survival (2,074 bulls) and overall type (1,407 bulls, 316 cows).

Estimation of SNP effects. Prediction equations to generate DGV were estimated from the bull data (reference set) by partial least squares regression (PLSR) and then used to predict DGV in cows (validation set). PLSR is a linear regression method that forms latent components as new independent predictors in a regression model (Moser et al. (2009)). The

[1] ReproGen, Fac. Vet. Sci. Univ. Sydney, 425 Werombi Rd. Camden, NSW 2570, Australia

[2] Biosciences Res. Division, DPI, 1 Park Drive, Bundoora, Vic 3083 Australia

components in PLSR are determined by both the response variable and the predictor variables. The magnitude of the PLSR regression coefficients ($ß_{PLS}$) can be used to determine which SNPs are most influential in the data set. To select subsets of markers all 42,567 SNP were ordered by their absolute value of $ß_{PLS}$. The SNP order was derived by a stepwise backward elimination procedure. The process starts with a model including the complete SNP set, subsequently in each step a fraction of SNP with the smallest $ß_{PLS}$ are dropped from the SNP list and the regression coefficients were recomputed. The optimal model complexity (i.e. number of latent components) was estimated by ten-fold cross-validation at each step.

SNP selection. Trait-specific subsets were selected by choosing the most influential SNP for each trait. To select subsets of SNP common across traits the SNPs with the largest $ß_{PLS}$ for ASI were chosen into the SNP panel. The $ß_{PLS}$ coefficients were then re-estimated for each trait to derive trait-specific prediction equations. Subsets of evenly spaced SNP were created by dividing the total length of the autosomes into intervals flanked by two markers of approximately equal length and then selecting the SNP with the highest minor allele frequency (MAF) or the largest $ß_{PLS}$ for ASI in each segment. SNP densities of 100, 300, 500, 1,000, 3,000 and 5,000 were compared with the full panel.

Criteria for comparison. The correlation coefficient between predicted DGV and the realized ABV of cows was used to evaluate the accuracy of predictions using subsets of SNP.

Results and discussion

Correlations between predicted DGV and realized ABV for cows achieved with trait-specific subsets are shown in Figure 1. Accuracies of DGV prediction of cows from the analysis using all 42,576 SNP were 0.69, 0.61, 0.50 and 0.35 for protein percentage, ASI, overall type and survival, respectively. Omitting the least influential SNP, using only the top 40,000 SNP increased the accuracy of prediction for three out of the four traits. As expected correlations decreased as the number of SNP in the subset decreased, but accuracies plateau over a large range with respect to the size of subsets and a drop is only noticeably for subsets of less than 5,000 SNP. There is a strong relationship between the accuracy of predictions and the heritability of the trait. Prediction for traits with high heritability like protein percentage (h^2=0.56) and ASI (h^2=0.26) achieved higher accuracy than overall type (h^2=0.18) and survival (h^2=0.03) with low heritability. To achieve similar accuracy for a trait with low heritability to that achieved for production traits, more records will be needed.

Trait-specific subsets of 300 SNP selected on the magnitude of $ß_{PLS}$ provided between 60% and 92% of the accuracy achieved with the high-density assay. However, no SNP was in common between all 4 subsets of 300 SNP, and only 37 SNP were selected across all traits in the trait-specific subsets of 5,000 SNP. Therefore, combining the most influential SNP for each trait onto a single chip or developing multiple low-density assays might not provide adequate reductions in genotyping costs. The cost of genotyping can be reduced if a single marker set is used across traits. The accuracy of DGV prediction of cows using common subsets of evenly spaced markers relative to the DGV accuracy obtained with all 42,576 SNP across all four traits is shown in Table 1.

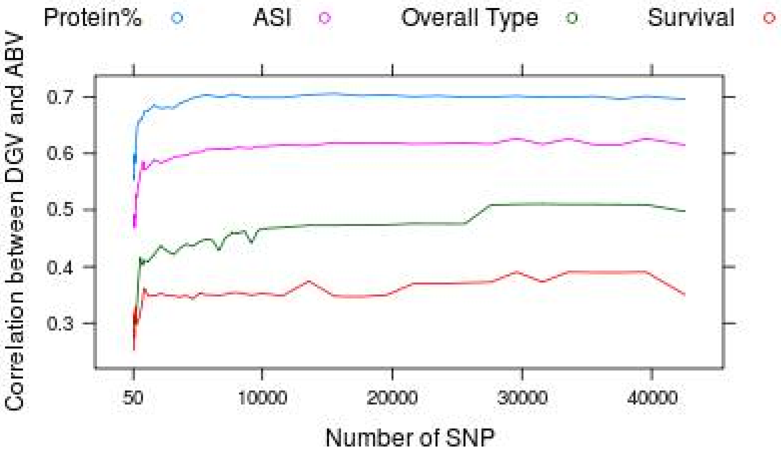

Figure 1: Correlation between DGV and ABV in cows using trait-specific subsets of SNP

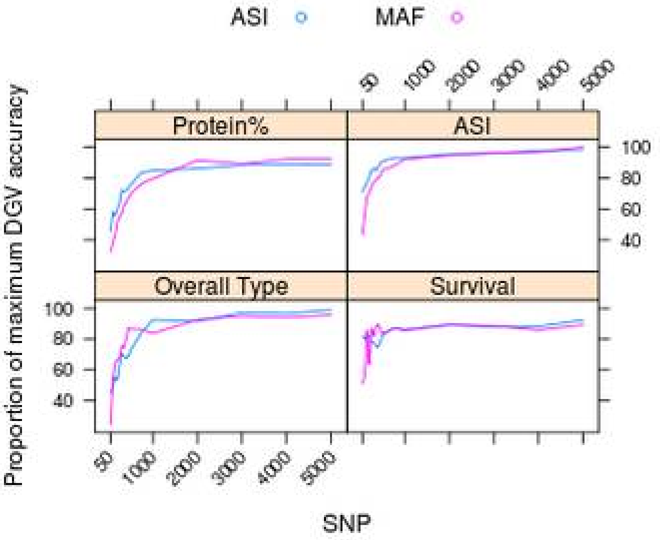

Figure 2: Proportion of the maximum DGV accuracy obtained with a common subset of equally spaced SNP selected for highest minor allele frequency (MAF) or their SNP rank for ASI

Table 1: Summary of subset selection for low-density cow chip using a common SNP set Shown is the accuracy of DGV prediction as a percentage of the accuracy of DGV obtained with 42,576 SNP, averaged over the four traits.

	Number of SNP					
SNP selection	**5,000**	**3,000**	**1,000**	**500**	**300**	**100**
Largest ß$_{PLS}$ for ASI	95	95	92	85	81	61
Largest ß$_{PLS}$ for ASI, evenly spaced	95	92	89	80	77	65
MAF, evenly spaced	94	92	85	81	73	51

For any given subset of SNP, accuracies of a subset of evenly spaced SNP selected for ASI were higher than for a subset selected for high MAF, but the differences were negligible for subsets = 3,000 SNP. It must be noted that the accuracy of DGV predictions is specific to the family structure of the data, with 92% of cows being daughters of sires in the reference set and cow records being included in the progeny information of their sires.

Conclusion

Accurate genomic evaluation of the Australian Holstein-Frisian cow population can be achieved with SNP chips containing ~ 1,000 to 5,000 SNP. Selecting 5,000 evenly spaced SNP gave 95% of the accuracy achieved with all 42,576 SNP, with none of the four traits having a relative accuracy less than 89%. An assay containing 1,000 evenly spaced markers can still provide over 85% of the accuracy obtained with the high-density SNP assay. A common marker set across all traits is preferred as weights can be updated and additional traits predicted from this common panel. It also allows for a high-volume generic chip to be produced which will lower assay cost per individual and is easier to implement in practice than multiple assays for different traits.

Acknowlegement

The research was funded by the Co-operative Research Centre for Innovative Dairy Products (CRC-IDP). The core data resource was generated by the CRC-IDP and the Department of Primary Industries, Victoria, with input from Drs, Phil Bowman, Amanda Chamberlain, Julie Cavanagh, Natasha Ellis, Kyall Zenger, and Mrs Gina Attard. The semen samples were donated by Genetics Australia.

References

Moser, G., et al. (2009). *Genet. Sel. Evol*., 41:56

Schaeffer, L.R. (2006). *J. Anim. Breed. Genet*., 123:218–223

Weigel, K.A., de los Campos, G., González-Recio, O., et al. (2009). *J. Dairy Sci*., 92:5248–5257