The value of combining high-density SNP sets on the accuracy of direct genomic values in Holstein bulls

H.W. Raadsma[1], G. Moser[1] and M.S. Khatkar[1]

Introduction

Efforts of large scale genome sequencing projects have lead to the discovery of many millions of Single Nucleotide Polymorphisms (SNPs) in bovine (Gibbs et al. (2009)). Parallel to the SNP discovery efforts, the development of generic high-density SNP genotyping platforms (Elsik et al. (2009)) has lead to a number of genome-wide high-density SNP marker panels or so called SNP chips. Two major providers of high-density genotyping arrays dominate the bovine area with both Affymetrix (http://www.affymetrix.com) and Illumina (http://www.illumina.com) providing bovine specific SNP sets. Typically such SNP sets contain SNP from private and public domain sources and generally provide little overlap with common SNP across chips. Access to high-density SNP panels has opened the possibility of prediction genetic merit in cattle based on genome wide SNP markers through estimation of direct genomic values (DGV) or molecular breeding values, using genomic selection (GS) methodologies (Meuwissen et al. (2001)). Typically GS consists of estimation of prediction equations for DGV from a reference population (training data) where animals are genotyped and phenotyped and validating these in an independent but related population (test set) and then applying these equations in genotyped individuals for which *denovo* predictions are sought. In dairy cattle typically such training and test sets are bulls with estimated breeding values (EBV) based on progeny performance data, and the predictions are applied in young candidates under selection before progeny testing (Schaeffer (2006)). The problem faced by industry is to make decisions on which genotyping platform to commit to, the density of genetic markers required, and if combining markers from diverse SNP panels will increase accuracy of prediction of DGV. In this study we use a relatively low-density SNP panel, and the highest density SNP panel currently available for bovine, and the combination of the two to asses their impact on prediction of DGV in young dairy bulls.

Material and methods

Phenotype and genotype data. Given that accuracy of genomic selection may be influenced by the heritability of the trait, reliability of the EBV for bulls in the training and test sets, and nature of the trait (a single biological measure or an index derived from multiple traits), we examined the impact of varying SNP density on four traits commonly considered in dairy cattle breeding, namely protein percentage, Australian Selection index (ASI) as a profit

[1] The Co-operative Research Centre for Innovative Dairy Products, The University of Sydney, NSW, Australia

index, survival index and overall type. The summary characteristics of each of the four traits in terms of reliability of EBV of bulls in training and test set, and heritability of the traits is shown in Table 1.

Table 1. Summary of number of bulls with phenotypes (N), reliability of EBV (%) and heritability (h^2) for each trait.

		Training set		Test set	
Trait	h^2	N	Reliability	N	Reliability
Protein percentage	0.56	1445	91.7	361	86.2
ASI	0.26	1445	91.7	361	86.2
Survival	0.03	1448	72.4	361	55.4
Overall type	0.18	1116	66.3	356	42.1

The phenotypes were deregressed Australian Breeding Values (ABV) for protein percentage, ASI 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. The ASI is given by (3.8 × protein ABV) + (0.9 × fat ABV) – (0.048 × milk ABV), whereas survival is given by (0.5 × likeability) + (1.8 × overall type) + (3.0 × udder depth) + (2.2 × pin set).

SNP genotypes were derived from the Affymetrix 15K panel as described by (Khatkar et al. (2007)) and an Affymtrix bovine 25K panel (Affymetrix; http://www.affymetrix.com), The two Affymetrix panels had 10,410 SNP in common of which after quality control a total of 7,372 SNP remained. The high- density genotypes were provided by use of the Illumina BovineSNP50 BeadChip. A total of 42,111 markers remained for the analysis from the Illumina SNP50, panel, which gave a combined data set of 47,090 SNP after accounting for SNP in common between the two data sets. Bulls in the training and test sets were genotyped with both the Affymetrix and Illumina panels.

SNP imputation*:* Applying imputation procedures, a total of 56,551 SNP data points were available for all animals in the data set. After comparing different imputation methods, we used Beagle (Browning and Browning (2007)). A number of test sets were created by masking all the known genotypes on the 15K or 25K chip for various proportions of animals and accuracies of imputed genotypes were between 96-98 % (Khatkar pers. Communication).

.

Estimation of DGV predictions. Prediction equations to generate DGV were estimated from older bulls born between 1955 and 2000 (training set) by partial least squares regression (PLSR, Moser et al. (2009)) and then used to predict DGV in 361 younger bulls born between 2001 and 2003 (test data). Marker panels of 7,372; 42,111; 47,090 and 56,551 SNP were used in the analyses.

Criteria for comparison. The correlation coefficient between predicted DGV and the realized EBV of bulls in test set was used as a measure of the accuracy of GS, derived from SNP panels with different SNP density.

Results and discussion

Correlations between predicted DGV and realized EBV for the bulls in the test set are shown in Table 2. Accuracies of DGV prediction using the lower density SNP panel of 7,372 SNP were 0.57, 0.38, 0.47 and 0.34 for protein percentage, ASI, overall type and survival, respectively. Using the higher density IlluminaSNP50 panel with 42,111 SNP, resulted in an average increase of accuracy of 0.058 (approximately 10%) across all four traits (0.65, 0.47, 0.50 and 0.39 for protein percentage, ASI, overall type and survival, respectively). Combining SNP information from both SNP sets resulted in 47,090SNP (2,755SNP were in common). The combined SNP set did not result in additional gain of the average accuracy of DGV prediction of the test set animals when compared to the use of the high-density Illumina SNP50 panel with 42,111 SNP. Similarly, imputing SNP genotypes from the medium density SNP chips increased the number of SNP available to 56,551. This currently represents the highest density of SNP available on commercial public domain SNP typing platforms in bovine. The highest SNP density did not result in an improvement of accuracy of DGV prediction over the use of the Illumina SNP 50 set with 42,111 SNP (Table 2).

Table 2. Accuracy of prediction of DGV at four densities of genome-wide SNP panels for four traits in a test set of approximately 360 Holstein Friesian bulls

Trait	7,372 SNP	42,111 SNP	47,090 SNP	56,551 SNP
Protein %	0.570	0.645	0.647	0.656
ASI	0.383	0.466	0.460	0.453
Survival	0.345	0.393	0.394	0.391
Overall type	0.469	0.495	0.496	0.496
Average	**0.442**	**0.500**	**0.500**	**0.499**

The accuracy of DGV prediction was highest for protein percentage which had the highest heritability and reliability of bulls in the training and test set, and ranged from 0.57 to 0.66 depending on the number of SNP in the analysis. The next highest accuracy of DGV prediction was for overall type (range 0.47-0.50), a trait with moderate heritability, but relatively low reliability of bulls in both the training and test set (see Table 1). Accuracy of DGV for the profit index ASI was lower than for protein percentage or overall type (range 0.38-0.47), despite bulls in the training and tests set having the same high reliability of EBV as for protein percentage. The lowest accuracy of DGV prediction was calculated for survival (range 0.34-0.39), a trait with a low heritability and low reliability of EBV. The trend in accuracy of DGV across the four densities of SNP genotyping scenarios was very consistent, despite the range of heritabilities and accuracies (reliabilities?) of EBV, in that predictions of all traits improved marginally by increasing the number of SNP from 7,372 to 42,111, with no further improvements in accuracy when moving to 47,090 and 56,551 SNP.

The results show that there is little advantage in genotyping animals with multiple SNP panels in order to increase the SNP density and genome coverage from 42k to either 47k or 56K. Although, these are not considered ultra-high density SNP panels by current standards, the increase in density may not have resulted in providing more SNP in higher LD with putative QTN. The development of ultra-high SNP panels – ie 1,000K (SNP panels will result in higher LD between SNP and QTN and may as such lead to improvements of accuracy of prediction of DGV. Similarly ultra-high density SNP panels may also lead to inclusion of rare variants. With the advent of ultra-high density SNP panels and evolution of low-cost next generation sequencing the value of imputation is likely to reside in accurately predicting whole genome content for animals in the training and test sets by sequencing key ancestors. It is unknown what minimum SNP density is required to impute whole genome content and track rare variants and structural variants not currently genotyped directly, such as copy number variants or small and large insertions and deletions.

Conclusion

At current densities of commercial SNP panels the genome wide SNP coverage appears to be sufficient to ensure accurate prediction of the next generation of selection candidates, if animals in the training and validation sets are more or less closely related (Habier et al. (2007)). Higher SNP densities are required to accurately predict DGV for 'unrelated' animals. To take advantage of high marker densities, large training data sets are needed (Meuwissen, (2009))

Acknowledgements

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 by the Department of Primary Industries, Victoria, with input from Drs, Phil Bowman, Amanda Chamberlain, Julie Cavanagh, Natasha Ellis, Ben Hayes, Kyall Zenger, and mrs Gina Attard. The semen samples were donated by Genetics Australia.

References

Browning S.R. and Browning B.L. (2007). *Am J Hum Genet* 81: 1084-97.
Elsik C.G., Tellam R.L., Worley K.C., *et al.* (2009). *Science* 324: 522-8.
Gibbs R.A., Taylor J.F., Van Tassell C.P., *et al.* (2009). *Science* 324: 528-32.
Habier D., Fernando R.L. and Dekkers J.C. (2007). *Genetics* 177: 2389-97.
Khatkar M.S., Zenger K.R., Hobbs M., *et al.* (2007). *Genetics* 176: 763-72.
Meuwissen T.H. (2009). *Genet Sel Evol* 41: 35.
Meuwissen T.H., Hayes B.J. and Goddard M.E. (2001). *Genetics* 157: 1819-29.
Moser G., Tier B., Crump R.E., *et al.* (2009). *Genet Sel Evol* 41: 56.
Schaeffer L.R. (2006). *J Anim Breed Genet* 123: 218-23.