Experiences with the Illumina High Density Bovine BeadChip

B. L. Harris, F. E. Creagh, A. M. Winkelman and D. L. Johnson
LIC, Private Bag 3016, Hamilton, 3240, New Zealand

Abstract

The effect of increasing SNP density on the accuracy and inflation of genomic predictions of protein yield was investigated. Three SNP densities were studied. One was based on the 50K chip (38,296 SNPs). The two others were based on the high-density (HD) chips; HD1 (692,598 SNPs) and a subset, HD2 (329,329 SNPs), obtained by reducing the multicollinearity of the HD1 SNPs. The data set consisted of 4211 Holstein Friesian (HF), Jersey (J) and HFxJ sires genotyped with the 50K chip. The HD1 data was obtained through either genotyping or imputation. The test animals (N=605) were the two youngest cohorts of the progeny-tested sires. The genomic breeding values (GBVs) of the test animals were predicted using Bayesian and nonparametric methods. All models included the polygenic effect.

A higher SNP density was found to slightly improve the accuracy of prediction when Bayesian ridge regression was used to train on one breed and predict on a different breed. When using an across-breed analysis, the increase in SNP density did not increase the accuracy of predicting the GBVs. Furthermore, the genomic predictions were inflated compared to the model that included only the polygenic effect, with predictions obtained using the HD genotypes having greater inflation than those obtained from the 50K genotypes. The training data set may not have been large enough to take full advantage of the HD genotypes.

Introduction

Genomic selection involves selection on genomic breeding values (GBVs) predicted from single nucleotide polymorphism (SNP) markers that cover the whole genome. Prior to 2011, genomic selection in many dairy cattle populations was based on the Illumina BovineSNP50™ beadchip or similar-sized SNP chip technologies. The Illumina BovineSNP50™ beadchip technology contains 40-50,000 SNP markers. In 2011, the Illumina HD (high density) Bovine beadchip became commercially available. The Illumina HD Bovine BeadChip contains 770,000 markers. One of the goals of moving to the HD chip was to enhance the accuracy of genomic predictions within and across breeds. The increase in accuracy is assumed to result from a reduction in the physical distance between the SNP marker(s) and the QTL, thereby leading to greater levels of linkage disequilibrium (LD) that is preserved across generations within breed and across breeds.

A number of simulation studies have been undertaken to investigate the effect of SNP marker density on the accuracy of genomic evaluation (Meuwisson and Goddard, 2010, Harris and Johnson, 2010 and VanRaden, 2010). The results from these studies vary due to the initial assumptions about underlying distribution of QTL effects, the statistical methods employed and the amount of LD simulated between the QTL and SNP markers. Meuwisson and Goddard (2010) found that the accuracy of genomic selection increased approximately linearly with the log of the number of SNPs, whereas the other studies found little increase in the accuracy with increasing SNP density. Meuwissen and Goddard (2010) showed that in situations where there were fewer QTL, predictions from methods such as Bayes B provided increased accuracy as the SNP density increased in modest training population sizes.

De Roos *et al.* (2009), Ibanez-Esciche *et al.* (2009), and Pryce *et al.* (2011) have shown that across-breed genomic evaluations are more accurate than within-breed evaluations. The closer the genetic distance between the breeds and the greater marker density, the greater the improvement in accuracy.

The first aim of this study was to assess the accuracy of genomic prediction across breeds, using both the Illumina BovineSNP50 (referred to as 50K) and HD chips. The second aim was to compare the accuracy and inflation of genomic predictions obtained from different statistical models using an across-breed analysis and genomic information from either of the two chips. In all cases, the trait modelled was protein yield.

Methods

Data

Daughter yield deviations (DYD) for first-lactation protein yield were available on 4211 progeny-tested sires. The accuracies of the DYDs were used to calculate the weights according to the method described by Garrick *et al.* (2009). Three breeds were represented in the data: Holstein Friesian (HF), Jersey (J) and Holstein Friesian-Jersey HF-J crossbreds. To assess the accuracy of genomic prediction across breed, the sires were divided into 3 combinations of training and test data, as shown in Table 1.

The data set was split into training and test sets based on birth year. The test data contained 605 sires from the two youngest cohorts of the progeny-tested sires. The training data contained 3606 sires.

Table 1. The combinations of training and test data based on breed.

Training Data	Test Data
HF and Jersey Sires	HF-J Sires
HF Sires	Jersey and HF-J Sires
Jersey Sires	HF and HF-J Sires

HF-J = Holstein Friesian-Jersey crossbeed and HF = Holstein Friesian

All sires were genotyped on the 50K SNP chip (which contained 38,296 SNPs after quality control). A subset of the sires was also genotyped on the HD SNP chip, the remaining sires received HD genotypes via imputation. The methods used and the accuracies obtained from imputation are given in Johnson *et al.* (2011). Two data sets containing the HD genotypes were created after SNP quality checks were completed. The first data set (HD1) contained 692,598 SNPs. The second data set (HD2) was obtained by removing one of a pair of SNPs within a 250-SNP interval on the same chromosome that were in near perfect LD (R^2 >0.99), leaving 329,329 SNPs.

Statistical Analyses

Across-breed comparison

SNP effects were estimated in the training data sets using Bayesian ridge regression (BRR). The method assumed homogeneous shrinkage of SNP effects. A polygenic effect, with the variance structure described by the numerator relationship matrix, was included in the model. A Gibbs sampler similar to that described by Campos *et al.* (2009) was used to estimate the model effects, unknown variances and future observations for individuals in the test data. The accuracy of prediction was estimated as the correlation between the observed and estimated future records for the animals in the test data.

Accuracy of genomic selection

Three different statistical analyses were undertaken to estimate SNP effects and estimate variances. All the analyses included the polygenic effect. The first analysis used was a BRR described above. The second analysis used a Bayesian Lasso (BL) (Campos *et al.,* 2009). The BL was solved using the BLR package in R (Perez *et al.,* 2010). The BL model was not applied to the HD1 data due to computational limitations. The third analysis used reproducing kernel Hilbert

space (RKHS) regression methods. Two RKHS models were investigated. The first fitted all the SNP using a Gaussian kernel (Gianola and van Kaam *et al.*, 2008) with a single smoothing parameter while the second one used a smoothing parameter for each chromosome. The RHKS models were solved using a Gibbs sampler based on the Bayesian framework outlined by Gianola and van Kaam (2008). The Gibbs sampler for all models was run for 100,000 iterations, with the first 20,000 iterations used as burn-in. After burn-in, samples from every 10^{th} iteration were collected thereby avoiding autocorrelation among consecutive samples. The accuracy of prediction was calculated as the correlation between the observed and estimated future records for the test animals. The inflation was assessed using the regression slope of the observed on estimated future records, a slope of unity indicating no inflation.

Results and Discussion

Across-breed comparison

A major advantage of increasing SNP density in genomic selection is thought to be the increase in LD between flanking SNP markers and the QTL, thereby increasing the power of the markers. Table 2 shows the accuracy of predicting first-lactation protein yield BVs across breeds, using the 50K and HD1 data. The increase in SNP density did not improve the accuracy of predicting the HF-J crossbreed BVs from the purebred BVs. This is to be expected since the SNP effects estimated in the crossbreed sires would be linear combinations of the SNP effects in the two parent breeds. There were moderate improvements in prediction accuracy using the HD1 data compared to the 50K data when predicting Jerseys from HFs, or vice-versa. The differences in the magnitudes of the accuracies are likely to be a function of the ratio of training and test data sizes.

Table 2. The accuracy of predicting first-lactation protein yield GBVs across breeds using the 50K and HD chips.

Analysis	Training data size	Test data size	Accuracy 50K	Accuracy HD1
HF and J to predict HF-J	3713	498	0.67	0.67
HF to predict J	2290	1423	0.37	0.41
J to predict HF	1423	2290	0.15	0.22

HF-J = Holstein Friesian-Jersey crossbeed and HF = Holstein Friesian

Table 3. The accuracy (correlation) and inflation (slope) of predicting first-lactation protein yield GBVs of the 2006 and 2007 progeny test sires evaluated using the 50K chip.

Analysis	Correlation	Slope
BRR	0.607	0.852
BL	0.618	0.962
RKHS	0.606	0.928

Table 4. The accuracy (correlation) and inflation (slope) of predicting first-lactation GBVs on the 2006 and 2007 progeny test sires evaluated using the HD chip.

	HD1 data		HD2 data	
Analysis	Correlation	Slope	Correlation	Slope
BRR	0.594	0.821	0.589	0.818
BL	n/a	n/a	0.600	0.872
RKHS	0.593	0.943	0.587	0.958
RKHS fitting each chromosome separately	n/a	n/a	0.591	0.870

Accuracy of genomic selection

The accuracy and inflation of genomic prediction for the three methods of analysis are given in Tables 3 and 4 for the 50K and HD data, respectively. The estimates are weighted averages of the within-breed estimates.

Estimates of the accuracy and inflation from a traditional polygenic model were obtained from a BRR analysis. The accuracy and inflation from this model was 0.531 and 0.894, respectively. In all cases, models that included the SNP effects resulted in higher accuracies compared with the polygenic model. However, inflation estimates from the models that included the SNP effects were consistently greater than that obtained from the polygenic model.

The accuracies of prediction using 50K data were higher than those obtained from either of the HD data sets. Within a given SNP density, the BL had marginally greater accuracy than the RKHS and BRR analyses. The BL and RKHS analyses had lower levels of inflation than the BRR analyses within a given a SNP data set.

The RKHS analysis that fitted each chromosome as a separate kernel matrix produced similar accuracy to RKHS analysis with a single kernel matrix for the SNP effects, but had greater inflation. The ratio of the variances obtained for each of chromosome kernel matrices, relative to the total SNP variance, was not proportional to the chromosome length. Chromosome 14, where the DGAT1 gene is located, had largest variance.

In all the analyses a polygenic effect was fitted. Excluding the polygenic effect resulted in lower accuracies and slightly greater inflation in all cases. The percentage of genetic variance attributed to the polygenic effect ranged from 2% to 7% across all analyses.

This study found that an increased SNP number resulted in a modest improvement in the ability to predict GBVs from one breed to another breed. However, no improvement was seen in the accuracy of prediction GBVs in the across-breed analysis. This study has presented results for protein yield only. However, similar results have been obtained for milk and fat yield, live-weight and somatic cell score.
A possible explanation for the lack of increase in accuracy with increasing SNP density may be that training population contained too few individuals. To make full use of the increase in the SNP densities, the number of animals in the training data may need to also increase.

References

Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. & Cotes, J.M. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics 182,* 375.

de Roos, A.P.W., Hayes, B.J., Spelman, R. & Goddard, M.E. 2009. Reliability of genomic breeding values across multiple populations. *Genetics 183,* 1545.

Garrick, D.J., Taylor, J.F. & Fernando, R.L. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol. 41,* 55.

Gianola, D. & van Kaam, J.B.C.H. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics 178,* 2289.

Harris, B.L. & Johnson, D.L. 2010. The impact of high density SNP chips on genomic evaluation in dairy cattle. *Proceedings of the 2010 Interbull Meeting,* Riga, Latvia. *Interbull Bulletin 42,* 40-43.

Ibánez-Escriche, N., Fernando, R.L., Toosi, A. & Dekkers, J.C.M. 2009. Genomic selection of purebreds for crossbred performance. *Genet. Sel. Evol. 41,* 12.

Johnson, D.L., Spelman, R.J., Hayr, M.K. & Keehan, M.D. 2011. Imputation of single nucleotide polymorphism genotypes in a crossbred dairy cattle population using a reference panel. *AAABG Conf. Proc. 2011,* Perth, Aus.

Meuwissen, T.H.E. & Goddard, M.E. 2010. Accurate pediction of genetic values for complex traits by whole genome resequencing. *Genetics 185,* 623.

Perez, P., de los Campos, G. & Gianola, D. 2010. Genomic enabled pediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *The Plant Genome 3,* 106.

Pryce, J.E., Gredler, B., Bolormaa, S., Bowman, P.J., Egger-Danner, C., Fuerst, C., Emmerling, R., Sölkner, J., Goddard, M.E. & Hayes, B.J. 2011. Genomic selection using a multi-breed, across-country reference population. *J. Dairy Sci. 94,* 2625.

VanRaden, P.M., O'Connell, J.R., Wiggans, G.R. & Weigel, K.A. 2010. Combining different marker densities in genomic evaluation. Proceedings of the 2010 Interbull Meeting, Riga, Latvia. *Interbull Bulletin 42,* 113-117.