The Impact of High Density SNP chips on Genomic Evaluation in Dairy Cattle

B. L. Harris and D. L. Johnson
LIC, Private Bag 3016, Hamilton, New Zealand

Abstract

New SNP chips for the bovine with densities of 500,000 to 800,000 SNP markers will become available for genomic evaluation of dairy cattle within the next 2 years. The aim of this study is evaluate the impact of increased SNP marker density on the accuracy of genomic evaluation using a simulation. This study has found that if the underlying genetic structure is defined by a large number of small QTL then large training data set sizes and precise phenotypic measures will be required to realise improvements in accuarcy from increasing SNP density.

Introduction

Genomic selection utilises breeding values that are predicted from a large number of estimated single nucleotide polymorphism (SNP) marker effects that cover the whole genome. The SNP marker information can be used to identify animals that have inherited chromosome segments of high genetic merit. Genomic selection allows juvenile animals to be selected with greater accuracy compared to traditional genetic evalution systems. The higher accuracy of selection can lead to increased genetic gain from the incorporation of genomic information in to dairy cattle breeding programs. Many dairy cattle populations contain animals, usually elite males and females, that have been genotyped using the Illumina BovineSNP50 BeadChip or similar technologies. The current SNP chip technology provides a density of 40-50,000 SNP markers for genomic analysis. A number of countries have incorporated the genomic information in their national genetic evaluations. The incorporation of genomic information has increased the accuracy of young bulls from a traditional parent average value of 35% to values greater than 50%. This increase in the young bull accuracy has resulted in increased use of young bulls to breed cows and changes to dairy cattle breeding scheme designs.

New SNP chips for the bovine with densities of 500,000 to 800,000 SNP markers are becoming available. It is likely that higher density SNP chip or genome sequence data will become available for the genomic evaluation of dairy cattle within the next 2 years. Analysis methods currently used for genomic evaluations in dairy cattle are largely based on ridge regression (RR) (VanRaden, 2008) which uses all the SNP markers. The success of this method in analysing dairy cattle data that uses all SNPs compared to methods that assume only a few SNPs are useful (Meuwissen *et al.,* 2001) raises questions about the underlying genetic structure. The success of ridge regression suggests that there may be a large number of smaller QTL. In this study the genetic structure simulated had a greater number of QTL with smaller effects compared to the studies reported by Meuwissen *et al.,* 2001 and Meuwissen and Goddard 2010.

The aim of this study is evaluate the impact of increased SNP marker density on the accuracy of genomic evaluation a simulation that is representative of the LIC breeding scheme in New Zealand. Also the limiting factors in terms of data and methods of analysis were determined.

Methods

Simulation

The SNP and QTL data were simulated to represent a dairy cattle breeding scheme using a simulation method similar to that used by Habier *et al.* (2007). A founder population was

simulated over a period of 1000 generations at an effective population size of 100. The genome consisted of 10 chromosomes, each of length 100 cM. Initially marker loci were equally spaced across the genome. The initial alleles were sampled from a Bernoulli distribution with probability 0.5. Starting from a segregating population, mutation rates were set at 0.5×10^{-3} for markers and where mutations switched to the other allele. A total of 1500 QTL were selected at random from among SNPs with minor allele frequency $\geq$ 5% leaving 1,070,562 SNP markers. Real pedigree data was based on the 5,769 bulls born up to 2008 and some bull dams that were genotyped in the LIC genomic selection programme. There was a total of 24,017 individuals in the pedigree, inclusive of the ancestors of genotyped individuals. The simulated founder population formed the base population for a gene dropping process through the pedigree. The QTL effect at each polymorphic locus was sampled from a gamma distribution with shape parameter 0.4. A true breeding value (BV) was generated for each individual based on the QTL effects and an initial phenotype was then obtained by adding in random error with the same variance such that heritability was 0.5. Subsequent phenotypes were simulated with heritabilities of 0.75 and 0.95. Higher heritabilities were achieved by estimating new random errors by reducing the random error variance. The test data comprised the 970 genotyped bulls born in 2007 and 2008. Lower density SNP data sets were created by randomly subsampling the full SNP data by chromosome. A set of 15,000 genotyped progeny test daughters were simulated with the sire chromosomes being randomly sampled from the sires born in 2006 and the dam chromosomes from the base population.

Statistical Analyses

Three methods of statistical analysis were undertaken on simulated data, RR (BLUP), Fast Bayes B (Meuwissen et al., 2009), and Elastic Net (Zou and Hastie, 2005). The Elastic Net (EN) method utilises the Lasso and ridge penalty to select variables. The EN has the interpretation as a stabilised version of the lasso. The EN has been shown to provide results that are comparable to Bayes B and RR for genomic selection using simulated and real data (Harris and Johnson, 2010). A multistep EN alogrithm was applied in this study. The first step was to select SNPs chromosome by chromosome. The second step was to analyse the selected SNPs aggregated across chromosomes. The multistep method has two advantages over a single-step analysis. First, the initial SNP selection is unaffected by correlations among SNPs on different chromosomes. Second, the multistep method is computat-ionally efficient compared to a single step analysis by factor of 3-4 times.

Results and Discussion

The major advantage of increasing SNP density in genomic selection is the improvement in linkage disequilibrium (LD) between flanking SNP markers and the QTL. Higher levels of LD provide a better QTL signal across and within families. Box plots of the LD between the 10 SNP markers flanking the QTL for 1000k and 20k SNP densities are provided in Figure 1. The greater the SNP density the higher the mean LD and the lower the range of LD between the QTL and the flanking markers. There are two disadvantages from increasing SNP density. First, the number of the uninformative SNPs in the data increases and linear functions of the uninformative SNPs may predict random error in the training phenotypes. This will reduce the accuracy of prediction observed in the test data. Second, the level of collinearity between uninformative and informative SNPs increases. The increasing collinearity can attribute the true QTL effect to a large number of correlated SNPs, thereby reducing the effectiveness of the prediction in future generations.

The correlations between the true and predicted phenotype in the test data for increasing SNP densities are given Table 1. There is little difference among the accuracies from the different analysis methods. The maximum correlation achievable in the test data is 0.71 when the heritability of the phenotype is 0.50. Only very small marginal changes in accuracy are seen, 0.63 to 0.66, as the SNP density increases from 20k to 1000k.

Table 1. The correlations between the true and predicted phenotype from the test data with increasing SNP density.

	SNP Density			
Analysis	20k	100k	500k	1000k
Elastic Net	0.643	0.645	0.646	0.651
RR[1]	0.650	0.653	0.653	0.655
FBB[2]	0.633	0.652	0.653	0.652

[1]Random Regression, [2]Fast Bayes B

To understand the drivers for the lack of improvement in the accuracy of prediction from increasing SNP density, three areas were investigated, the precision of phenotype, training data power and inclusion of the QTL SNPs in to the SNP data sets.

Table 2 provides the correlations between the true and predicted phenotype in the test data for three levels of phenotype precision and four SNP densities from an EN analysis. Increasing the phenotype precision improves the accuracy of prediction across all the SNP densities. However, there is little improvement in accuracy as the SNP density increases for a given level of phenotype precision. To determine whether the training data had sufficient statistical power to resolve higher SNP densities two analyses were undertaken. First the SNP density was further reduced below 20k. Second, the training data was increased by the addition of 15,000 progeny test daughters. The results from reducing the SNP density are given in Table 3 and results from increasing the training data given in Table 4.

Table 2. The correlations between the true and predicted phenotype from the test data for three phenotype heritabilities with increasing SNP density.

Phenotype h^2	SNP Density			
	20k	100k	500k	1000k
0.50	0.643	0.645	0.646	0.651
0.75	0.749	0.747	0.747	0.751
0.95	0.889	0.895	0.890	0.897

Increasing the SNP density progressively from 1k to 20k shows an asymptotic increase in the correlations between the true and predicted phenotype. Similar trends were also observed at higher levels of phenotype precision. It appears from this analysis that the statistical power of training data with 4799 sires is at maximum around 20k SNP density in this simulation. A density of 20k in this simulation equates to 65k for the bovine genome. The addition of 15,000 extra genotypes to the training data set improved the accuracy of prediction in the test data. The additional genotypes also appear to move the accuracy asymptote from 20k to approximately 100k (300k on the bovine genome).

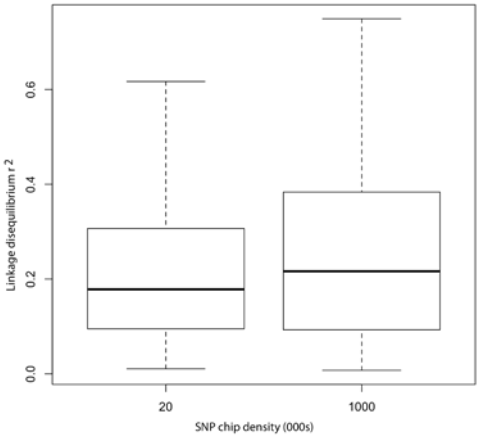

Figure 1. Box plot of average linkage disequilibuim r^2 for the 10 markers flanking the 1500 QTL.

To investigate the impact of including caustive mutations on the SNP chip, the QTL SNPs were included in SNP data. Table 5 provides the results from adding the QTL SNPs to the 20k SNP data for the three levels of phenotype precision. The correlations between the true and predicted phenotype increase with addition of the QTL SNPs. Further analysis of the addition of QTL SNPs showed that there was an observed interaction between the improvement from the addition of QTL SNPs, SNP density and phenotypic precision. As the SNP density increases and the phenotypic precision decreases the effectiveness of the addition of QTL SNPs

decreases. The signal from the QTL SNPs appears to be reduced by increased multicollinearity amongst SNPs and increased noise from the errors in the phenotypes.

Table 3. The correlations between the true and predicted phenotype from the test data with decreasing SNP density.

	SNP Density			
Analysis	1k	5k	10k	20k
Elastic Net	0.572	0.623	0.636	0.643
RR[1]	0.598	0.625	0.648	0.650
FBB[2]	0.512	0.576	0.616	0.633

[1]Random Regression, [2]Fast Bayes B

Table 4. The correlations between the true and predicted phenotype from the test data with additional training data genotypes

Training	SNP Density		
data	20k	50k	100k
4799 Sires	0.643	0.644	0.645
[1]4799+15k	0.661	0.678	0.693

[1]4799 Sires + 15k progeny test daughters

Table 5. The correlations between the true and predicted phenotype from the test data when the QTL SNPs are added to the 20k SNP data.

SNP	Phenotype h^2		
Density	0.50	0.75	0.95
20k	0.643	0.749	0.889
20k + QTL	0.666	0.811	0.937

The results in this study differ from those reported by Meuwissen and Goddard (2010) where increasing the SNP density improved the prediction accuracy in their simulations. The major difference between this study and the study by Meuwissen and Goddard's study is the simulation of the underlying genetic structure. In this study the number of QTL simulated was greater and the magnitude of the individual QTL effects smaller. Meuwissen and Goddard (2010) have shown that in situations where there a fewer QTL, predictions from methods such as Bayes B provide increased accuracy as the SNP density increases for modest training population size.

The increased accuracy of genomic selection prediction from increasing SNP density will be senstive to the underlying genetic structure. In situations where the underlying genetic structure is defined by a larger numbers of small QTL, large training data set sizes and precise phenotypic measures will be required to realise improvements in accuarcy from increasing SNP density.

References

Habier, D., Fernando, R.L. & Dekkers, J.C.M. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics 177,* 2389-2397.

Harris, B.L. & Johnson, D.L. 2010. SNP selection using elastic net, with application to genomic selection. *Ninth World Congress on Genetics applied to Livestock Production.*

Meuwissen, T.H.E., Hayes, B.H. & Goddard M.E. 2001. Accurate pediction of genetic values for complex traits by whole genome resequencing. *Genetics 157,* 1819-1829.

Meuwissen, T.H.E., Solberg, T.R., Shepherd, R. & Wooliams, J.A. 2009. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *GSE 41,* 2-11.

Meuwissen, T.H.E. & Goddard, M.E. 2010. Prediction of total genetic value using genome-wide dense marker maps. *Genetics 185.*

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91,* 4414-4423.

Zou, H. & Hastie, T. 2005. Regularization and variable selection via the elastic net. *J. Statist. Soc. B, 67,* 301-320.