Genomic selection in New Zealand and the implications for national genetic evaluation

B.L. Harris, D.L. Johnson & R.J. Spelman

LIC, Private Bag 3016, Hamilton, New Zealand

Summary

The BovineSNP50 BeadChip provides genome-wide dense-marker data. LIC genotyped 4,500 sires using the BovineSNP50 BeadChip. Genomic breeding values (GBVs) were estimated for 25 traits. There were several methods used to analyse the SNP data. Increased accuracy was obtained by using MCMC Bayesian methods compared to BLUP. However, the increased accuracy was small, often in the range of a 2-3% increase in the correlation between progeny test and genomic breeding values. The training statistical analysis was undertaken within and across breeds. The SNP estimates calculated from the Holstein-Friesian (HF) training data did not produce accurate GBVs in the Jersey test data and vice-versa. However, correlations for the crossbreed bulls (HF – Jersey crosses) were largest when all sires of all breeds were included in the training data subsets. GBVs for young sires without phenotype records from daughters had reliabilities of between 50-67% for the milk production traits, live weight, fertility, somatic cell and longevity. This increase in reliability compared to parent average has allowed LIC to reduce the size of the progeny test scheme, and to market teams of unproven bulls based on their GBVs. Methods to incorporate genomic information in to national and international genetic evaluation systems need to be developed. If genomic information is not integrated in to these systems, these systems will no longer be the primary source for selection information in the future.

***Key words**: Genomic evaluation, Single nucleotide polymorphism, Dairy cattle, Genomics, Selection, Markers, Breeding value.*

Introduction

Genome-wide dense marker-data are available for dairy cattle. Recently, Illumina, Inc. (San Diego, CA), have released the BovineSNP50 BeadChip that was developed by Van Tassell *et al.,* (2008). The BovineSNP50 BeadChip contains over 50 000 SNP uniformly distributed over the genome. Genome-wide SNP data can be used to estimate genomic breeding values (GBVs) as outlined by Meuwissen *et al.,* (2001). First, SNP effects are estimated with a training data set containing individuals with marker genotypes and trait phenotypes. Second, GBVs of any genotyped individual in the population can be calculated using the estimated marker effects. LIC genotyped 4 500 sires born in the 1980s to the present using the BovineSNP50 BeadChip. The data included approximately 2 400 Holstein Friesian, 1 500 Jersey and 650 Holstien Friesian–Jersey crossbred sires. The data was divided into training and test data sets based on birth year. Analyses were undertaken both within breed and across breed.

The test data sets were used to assess the accuracy of the GBVs by comparing BVs calculated from phenotypic records on daughters with GBVs calculated from SNP data. This paper will outline the results from the SNP analysis, the likely changes to the LIC progeny testing program and discuss the implications for the national genetic evaluation system.

Analysis of the genomic marker data

The SNP data on the 4 500 sires was subjected to a number of data checks. Approximately 1 400 loci had a deletion or an identified third allele which resulted in lower than average call rates and these SNPs were removed. SNPs were removed for low call rates, minor allele frequencies less than 2%, non-Mendelian inheritance and failed Hardy Weinberg tests. A total of 44 146 SNPs were retained for analysis from the original 48 465. The SNP data included 0.13% missing values. The missing values were imputed using breed and sire information. Also, the statistical analysis included an option to remove SNPs in complete or near complete linkage disequilibrium with SNPs on the same chromosome.

The first data analysis undertaken was a validation analysis. Genomic breeding values (BVs) were estimated for 25 traits; milk volume, milkfat, protein 270-day yields, live weight, fertility, somatic cell score, longevity, 12 linear conformation traits, 4 farmer scored linear traits, calving difficulty and gestation length. The input data for the analysis was de-regressed breeding values (Banos and Sigurdsson, 1996) with expected daughter equivalents (Fikse and Banos, 2001) as the associated measure of accuracy of each data point. The de-regression of breeding values avoided double accounting of genetic relationships in both the numerator relationship matrix and the SNP effects. The data were divided into two subsets, a training subset and a test subset. The training subset included all sires born prior to 2001. The test subset included all sires born between 2001 and 2003. Sires in both the training and test subsets had estimated BVs from progeny testing. The SNP effects were estimated in the training data set and then these estimated effects were used to estimate GBVs in the test data set. The accuracy of the GBVs was estimated by the correlation between these GBVs and the BVs based on progeny test daughters.

There were several methods of analysis used for training; BLUP, BayesA and BayesB (Meuwissen *et al.,* 2001), Bayesian regression (Xu *et al.,* 2003), Bayesian model selection (George and McCulloch, 1997), least angle regression (LARS) (Efron *et al.,* 2004) and Bayesian Inference Machine Learning (Tipping, 2001). On average the accuracy of LARS and Bayesian regression was considerably poorer than the other methods regardless of trait or breed. In a number of cases the BayesA, BayesB and Bayesian model selection methods outperformed BLUP. However, the increased accuracy obtained by these methods was small, often a 2-3% increase in the test correlations compared to BLUP.

The training statistical analysis was undertaken within and across breeds. The SNP estimates calculated from the Holstein-Friesian (HF) training data did not produce accurate GBVs in the Jersey test data and vice-versa. The test correlations ranged from -0.1 to 0.3 when the SNP effects from one breed were used to calculate GBVs in another breed. Conversely, the test correlations for the crossbreed bulls (HF – Jersey crosses) were largest when all sires of all breeds were included in the training data subsets. Overall the test correlations for the crossbreed bulls were 5-10% higher than for each of the two breeds.

Genomic breeding values for unproven sires

From the validation analysis it was apparent that Genomic BVs should calculated within breed for HF and Jersey sires and assigning equal variance to each SNP was close to optimal. To compute genomic BVs for proven and unproven sires simultaneously the genomic mixed model equations (MME) outlined by Van Raden (2007) and Habier *et al.,* (2007) was used. The input data for the models was de-regressed breeding values for proven sires. The MME uses a genomic relationship matrix rather than a numerator relationship matrix. In the MME analysis all the data is used, there is no distinction between training and test data subsets. The specification of genomic relationship matrix is flexible allowing individual SNP effect to have their own variance ratio. For this analysis the genomic relationship matrix was formed assuming equal variance for each SNP analogous to the BLUP method of (Meuwissen *et al.,* 2001). The reliabilities of GBVs were estimated by direct inversion of the mixed model equations. The GBVs for young bulls do not contain all the parent information that is contained in the national genetic evaluation. To incorporate the parent information from the national genetic evaluation a selection index approach was used. The parent information from the national genetic evaluation and the GBV are not independent sources of information. The covariance between the parent information and the GBV was approximated by the reliability of the additive parent information in the genomic MME. This was calculated substituting the genomic relationship matrix with the numerator relationship matrix based on sire-maternal-grandsire relationships in the genomic MME. Furthermore, the parent average and genomic BVs were expressed as deviations from a breed and birth year mean prior to applying the selection index to avoid over expansion of fixed breed and genetic groups effects contained with the BVs. The resulting breeding values for young sires had reliabilities of between 50-67% for the milk production traits, live weight, fertility, somatic cell and longevity compared to an average of 34% for parent average BVs. Reliability for the linear type traits were lower, ranging from 40-50%, compared to an average of 31% for parent average BVs.

Providing breeding values on young sires based solely on genomic and parent information highlights the problem of the overestimation of parent average as a predictor of breeding values based on progeny test for elite sires. The overestimation is commonly caused by (intentional or unintentional) preferential treatment and non-additive genetic effects in bull dams, such as epistasis and dominance, that are not passed on to offspring. It is important to remove the bias from the young sires breeding values.

LIC progeny testing program redesign

In a conventional dairy progeny testing system, it takes 7 years to obtain a proven bull from progeny testing. Currently, LIC progeny tests 300 bulls per year. Each bull obtains a first proof based on 80 daughters. The unproven bulls are used in approximately 500 dedicated sire proving scheme herds. Genomic selection using the 50k Illumina SNP chip can generate breeding values on unproven bulls with a reliability of 50-60% compared to 35% on parent information alone. This increase in reliability allows breeding companies to reduce the size of the progeny test scheme and to market teams of unproven bulls based on their combined SNP and parent average breeding values. LIC is currently redesigning its breeding scheme to market teams of young bulls at 2 years of age. The young bulls will be used in the sire proving scheme as yearlings to generate future phenotypic records from daughters and check that the young sires do no transmit any gross genetic abnormalities to their offspring. The size of sire proving scheme will be considerably smaller than the current scheme. The acquisition of young bulls to enter the progeny testing scheme is likely to change dramatically. In the current scheme 1500 elite females are contracted to produce 300 bull calves. In the future, it is possible that elite females and bull calves will be screened using the SNP chip technology. It is envisaged that thousands of bull calves will be screened prior to entry and the best 100-150 bull calves selected for progeny testing. The rate of genetic gain in the Breeding Worth index (a measure of net profitability) is likely to increase 50-60% to approximately 0.3 genetic standard deviation units per year.

Impact of genomic selection on the national genetic evaluation

The national genetic evaluation system calculates breeding values for cows and sires and is based on weighting three sources of information; parent information, own performance and progeny performance. Conceptually genomic selection provides a fourth source of information, SNP or DNA information.

Integrating genomic information in to an existing genetic evaluation system is more difficult than integrating other new information sources such as foreign daughter performance. This is because the accumulation of information does not follow the usual numerator relationship rules. For example, in conventional genetic evaluation, a sire with an increasing number of daughters lactating for the first time will see a steady increase in his reliability as the extra information accumulates. Also, diminishing returns prevail as the number of daughters increase. In a genomic setting, where the sire is genotyped, as an increasing number of his progeny are genotyped there is no extra information or increased reliability attributed to the sire. New approaches are required for combining genomic information with conventional breeding values. It may be possible to combine the numerator relationship matrix (for non-genotyped animals) and genomic relationship matrix (for genotyped animals) in a single analysis. In such an analysis it would be important to distinguish modes of SNP relatedness, identity by descent and identity in state, and that these modes are modeled appropriately.

Solving large-scale national genetic evaluation systems are currently feasible because simple rules exist to invert the numerator relationship matrix (Henderson 1976). Procedures such as iteration on data (Misztal and Gianola, 1987) are efficient for solving large-scale systems of equations in part due to the sparse nature of the inverse of the numerator relationship matrix. In the genomic MME, the genomic relationship matrix is a dense matrix. There are currently no rules to indirectly invert the matrix. As the number of genotyped animals increases with time, direct inversion of the genomic relationship matrix may become infeasible, and solving the genomic MME will become increasingly difficult.

Another challenge facing national genetic evaluation systems will be the emergence of SNP chips that contain a subset of SNPs from the BovineSNP50 BeadChip. These chips may allow large scale screening of individuals at a relatively low cost. This technology will enable breeding companies to preselect animals prior to being genotyped by the BovineSNP50 BeadChip. It is conceivable that breeding companies or individual dairy farmers would like the results from the smaller SNP panels to be incorporated in the national genetic evaluation system in the near future.

Interbull currently provides international comparisons among proven sires from a number of different countries. In New Zealand genomic evaluations are available on young sires with no progeny. Semen exporting companies will want Interbull international comparisons on young sires that include genomic information. Providing an international genetic evaluation system that combines daughter performance and genomic information from different sources will be a challenging project. The genomic information from different countries cannot be simply treated as independent sources of information and presumably environment by genomic interactions will exist as well. Countries using different combinations of SNPs and applying different weights to individual SNPs to predict future performance could further compound the problem.

Discussion

The availability of genome-wide dense marker data for dairy cattle has allowed GBVs to estimated on young sires with no progeny information. The GBVs are of sufficient accuracy to allow LIC to redesign its breeding scheme and market teams of young sires based on their GBV evaluations. Farmers will want genomic information incorporated into the national genetic evaluation when the first progeny from the young sires are born (July 2009). It is likely that the methods for calculating GBVs from SNP data will improve and the accuracy of GBVs will increase with time. As the accuracy of GBVs increase there will be a greater usage of young sires based on their GBV evaluations increasing the pressure on national genetic evaluation procedures to incorporate genomic information. Similarly, the semen exporting companies will require international genetic evaluations to include genomic information in the near future. If genomic information is not integrated in to these systems then these systems will no longer be the primary source for selection information and their usefulness is limited.

Technology will improve, higher density SNP chips will become available and in time complete sequence data could be available on individual sires at relatively low cost. The genetic evaluation solutions will need to evolve with the advances in DNA technology.

Acknowledgements

The authors thank Drs. George Wiggans and Paul VanRaden (USDA) for stimulating discussions during the authors visit to USDA Beltsville; and funding support from Dairy New Zealand.

List of references

Banos, G. & A. Sigurdsson. 1996 Application of contemporary methods for the use of international data in national genetic evaluations. J. Dairy Sci. 79(6): 1117-1125.

Efron B., T. Hastie, I.Johnstone & R. Tibshirani. 2004. Least Angle Regression. Annals of Statistics 32: 407-499.

Fikse, W.F. & G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. J. Dairy Sci. 84:1759–1767.

George E.I. & R.E. McCulloch. 1997. Approaches for Bayesian Variable Selection. Statistica Sinica 7: 339-74.

Habier, D., R.L. Fernando & J.C.M Dekkers. 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177: 2389-2397

Henderson C.R. 1977. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics, 32: 69–83.

Meuwissen, T.H.E., B.J. Hayes & M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Misztal, I. & D. Gianola. 1987. Indirect solution of mixed model equations. J. Dairy Sci. 70: 716–723.

Tipping, M.E. 2001. Sparse Bayesian Learning and the Relevance Vector Machine. J. of Machine Learning Research, 1: 211-244

VanRaden, P.M. 2007. Genomic measures of relationship and inbreeding. Interbull Bull. 37: 33–36.

Van Tassell C.P., T.P.L. Smith, L.K. Matukumalli, J.F. Taylor, R.D. Schnabel, C.T. Lawley, C.D. Haudenschild, S.S. Moore, W.C. Warren & T.S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5: 247–252.

Xu, Shizhong. 2003. Estimating Polygenic Effects Using Markers of the Entire Genome. Genetics 163: 789-801