National Genomic Evaluations without Genotypes

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

Introduction

Traditionally, performance and pedigree information are combined to estimate breeding values in a statistical framework, usually BLUP. The degree of identity by descent assumed in the statistical framework is based on probabilities derived from relationships contained in the known pedigree structure. An additional source of information based on DNA data has become available for national genetic evaluation systems. The DNA information facilitates tracing the inheritance of individual small segments of the chromosomes, thereby providing more identity by descent information than can be derived from the pedigree information alone.

The genetic similarity among individuals is based on thousands of markers spread across the genome and can be measured more accurately than similarity based on pedigree relationships (Meuwissen, 2007). Single nucleotide polymorphism (SNP) markers now cover the bovine genome with high density. Genomic predictions can be based on a BLUP-GS model where the average relationship matrix based on pedigree is replaced in the traditional BLUP model by a genomic relationship matrix based on markers (Habier *et al.*, 2007).

The DNA information leads to an increase in the accuracy of prediction of genomic breeding values, a decrease in generation intervals, and facilitates selection at a young age (Meuwissen *et al.*, 2001).

The objectives of this research were to apply genomic prediction methods to a population of Holstein Friesian, Jersey and Friesian x Jersey crossbred bulls and to integrate this information into the New Zealand genetic evaluation system for dairy cattle. An unusual feature of the integration of genomic data into the national genetic evaluation system is that the national centre does not have access to the raw SNP genotypes. The realized predictive ability of the genomic evaluations for bulls is also reported.

Linear Model and Genomic Relationships

The linear model relating phenotype **y** to SNP marker effects **u** is:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

where **Xb** represents fixed effects and e is random error with diagonal variance matrix **R**. The matrix **Z** has the mth column vector corresponding to SNP marker m and is coded -1, 0 and 1 for homozygote, heterozygote and other homozygote respectively. The sum over all SNP loci is assumed to equal the vector of breeding values (BV), **a**=**Zu**. Assuming fixed effects are known, the BVs can be estimated from:

$$\hat{\mathbf{a}} = \mathbf{Z}\hat{\mathbf{u}} = \mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{Xb})$$

where **D**=var(**u**) is a diagonal matrix of SNP variances. Matrix manipulation yields:

$$\hat{\mathbf{a}} = \mathbf{ZDZ}'(\mathbf{ZDZ}' + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{Xb})$$

and this can be further rearranged as:

$$\hat{\mathbf{a}} = (\mathbf{R}^{-1} + \sigma_u^{-2}(\mathbf{ZZ}')^{-1})^{-1}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{Xb})$$

where σ_u^2 is the common SNP variance in the BLUP approach.

Diagonals of ZZ' count the number of homozygous loci for each individual and off-diagonals (added to the number of SNPs) measure the number of alleles shared by two individuals. Based on the above expectation, a genomic relationship matrix can be estimated by a regression technique (VanRaden 2007) using the model:

$$\mathbf{ZZ'} = \sum_{k \le l} b_{1(kl)} \mathbf{X1}_{(kl)} + \sum_{k \le l} b_{2(kl)} \mathbf{X2}_{(kl)} + \mathbf{E}$$

where **E** includes differences between true and expected fractions of DNA in common as well as measurement error due to using a subset of the full DNA sequence. This regression method does not require estimates of allele frequencies (which should reflect founder population values) and was at least as good as, if not better than, two other methods discussed by VanRaden (2007). The genomic relationship matrix G can be estimated by

$$\mathbf{G} = \frac{\hat{b}_2}{\mathbf{ZZ'} - \hat{b}_1 \mathbf{X1}_{(kl)}}$$

Replacing regression coefficients by expected values then:

$$\hat{\mathbf{a}} = (\mathbf{R}^{-1} + \sigma_a^{-2}\mathbf{G}^{-1})^{-1}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{Xb})$$

where

$$\sigma_a^2 = 2\sum_m p_m q_m \sigma_u^2$$

is additive genetic variance, where p_m is allele frequency and q_m =(1–p_m). A solution strategy which avoids inverting G is:

$$\hat{\mathbf{a}} = \mathbf{G}(\mathbf{G} + \sigma_a^{-2}\mathbf{R})^{-1}(\mathbf{y} - \mathbf{Xb})$$

with reliabilities obtained for individual i as:

$$\frac{(\mathbf{G}(\mathbf{G} + \sigma_a^{-2}\mathbf{R})^{-1}\mathbf{G})_{ii}}{\mathbf{G}_{ii}}$$

Genomic relationship matrix for a multibreed population

The covariance between relatives in a multibreed population should take account of differences in allele frequencies among breeds. The regression technique used to calculate the genomic relationship matrix can be generalized to multiple regression. The regression technique is generalized to:

$$\mathbf{ZZ'} = \sum_{k \le l} b_{1(kl)} \mathbf{X1}_{(kl)} + \sum_{k \le l} b_{2(kl)} \mathbf{X2}_{(kl)} + \mathbf{E}$$

where $\mathbf{X1}_{(kl)}$ are the covariates for the means (intercepts) and $\mathbf{X2}_{(kl)}$ are the covariates for the regression components associated with covariances determined by the rules set down in Lo *et al.* (1993). This algorithm is similar to that for forming the relationship matrix in a purebred population except that when forming the diagonals we partition into breed fractions to account for different variances among breeds and include segregation variances due to different allele frequencies among breeds. The pedigree used for the construction of the above relationship matrices will inevitably contain some ungenotyped ancestors. The regression coefficients were estimated using the subset of genotyped animals. The generalization for estimation of the genomic relationship matrix for a multibreed population is then given by:

$$\mathbf{G} = \mathbf{L}_1 \hat{\mathbf{F}}_1^{-1} (\mathbf{ZZ'} - \sum_{k \le l} \hat{b}_{1(kl)} \mathbf{X1}_{(kl)}) \hat{\mathbf{F}}_1'^{-1} \mathbf{L}_1'$$

where $\mathbf{L}^{-1}$ and $\mathbf{F}^{-1}$ are derived from a Cholesky factorisation of $\mathbf{X2}_{(kl)}$ (see Harris and Johnson 2009).

Blended Genomic Breeding Values

The BV derived from the national genetic evaluation is made up of two components; a contribution from genetic groups (including breed effects) and a random additive genetic component. The phenotype used for the genomic predictions is the deregressed BV from the national genetic evaluation. Given the deregressed BVs, the genomic BVs (for the subset of genotyped animals) can then be estimated by the reverse process but replacing the numerator relationship matrix by the genomic relationship matrix. The genomic BVs may not contain all the information contained in the national BVs. There may be a loss of some of the parent-average information in the national evaluation. In order to recapture the information lost from non-genotyped animals a selection index is used to combine three sources of information for genotyped animals (Harris and Johnson 2009) and an updated reliability is calculated.

Integration in National Genetic Evaluation

The process for passing the genomic information through to ungenotyped individuals is only undertaken for descendants of genotyped individuals within a selection index procedure. We work down the pedigree

updating the BVs of ungenotyped individuals where either parent has a BV which has been updated with genomic information. If one or both parents has a genomic BV, then the descendant's breeding value can be updated based on the incremental genomic information. An updated reliability for the descendant is also calculated.

National Genetic Evaluation without SNP Data

The national genetic evaluation centre does not have access to the raw SNP data from the breeding companies and is unlikely to have access in the near to medium term. Furthermore, there are at least two SNP chips being used to genotype New Zealand based animals. Given these constraints the national genetic evaluation with genomic information has been developed to use genomic relationship coefficients derived from the raw SNP genotypes.

The process allows individual breeding companies to calculate the genomic relationship matrix for the animals that they have genotyped. The process involves three steps:

1. A computer program is run by the breeding company, conducting a series of checks on the raw SNP data. The program checks for monomorphic SNPs, sex chromosome SNPs, identical twins and SNPs in near perfect collinearity. The results of the program are reported to the national genetic evaluation centre.
2. A second computer program is provided to the breeding companies to calculate the genomic relationships. The output of this program is sent to the national genetic evaluation centre for inclusion in the genetic evaluation run.
3. On receipt of the genomic relationship matrix from a given breeding company a check on the individual coefficients is undertaken. The genomic relationship matrix is compared to the previous version to help eliminate potential processing errors. The overall measures of genetic variance and SNP allelic frequency are also compared.

The genomic breeding values are calculated for each genomic relationship matrix separately. This ensures that the genomic information from an individual breeding company does not have an impact on genomic breeding values from a different company. This approach is non-optimal from a total industry perspective compared to full sharing of all the genotypes. However, this allows different breeding companies to participate in a national genomic evaluation without the need to disclose and share genotypes.

A selection index procedure is used to combine genomic breeding values for animals which have multiple genomic evaluations arising from being genotyped by more than one breeding company.

Offical Genomic Breeding Value Release July 2009

National genetic evaluations including genomic information were released for all traits in July 2009. One breeding company, LIC, provided a genomic relationship matrix containing 4996 genotyped animals based on 42,306 SNPs on the Illumina BovineSNP50 BeadChip platform. The proven bulls on average had reliabilities increased 1-4% with genomic information. For unproven bulls, the national evaluation parent average of 35% reliability increased on average to 54% with genomic information and further to 57% with blending. The reliability increases were lower for fertility, somatic cell, linear type and longevity.

Historical Validation on the Genomic Data

A study was also undertaken to compare the genomic breeding values with traditional BVs in the national genetic evaluation system. National BVs were generated at the end of each season commencing spring 2000 through to spring 2008. Blended genomic BVs were computed for seasons 2000 to 2007 using genomic relationship matrices, pedigree and performance data available for the population at each point of time. The size of the genotyped reference population was increasing with time. Proven sires received BVs based on

parent, progeny and genomic information and young sires received BVs based on parents and genomics.

The genomic data used for the validation was provided by LIC. The genomic data contained 5212 genotyped animals, born since 1980, using the Illumina BovineSNP50 SNP panel – 2711 Holstein-Friesian (HF), 1738 Jersey (JE) and 763 Friesian-Jersey crossbred (FJ) bulls.

The top 15 unproven young bulls at 4 years of age and within breed were selected based on their parent-average BV (PABV) and based on their blended genomic BV (GBV). The number of bulls within these two groups remaining in the top 15 based in the 2008 evaluation, including daughter information, was recorded. Also the average protein BV of the two groups at the 2008 evaluation were calculated. Finally the coefficient of determination, the squared correlation between PABV or GBV at 4 years of age and the progeny-test BV (PTBV) at 5 years of age was calculated. These values were also calculated for the PTBV calculated at the 2008 evaluation.

Results from the validation showed that of 15 young bulls selected on parent average, an average of 6.4 were in the top 15 after daughter information was included. For selection on GBV this figure increased to 8.7. The average increase in protein BV (2008 evaluation) due to selecting the team on GBV compared to PABV was 2.1 kg. Genomic predictions increased coefficients of determination for all production traits that were studied. Coefficients of determination were 14 to 20 % higher for GBV relative to PABV when comparing to first PTBV and 4 to 26% higher when comparing to the PTBV at the 2008 evaluation.

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. 2009. Genomic Predictions for New Zealand Dairy Bulls and Integration with National Genetic Evaluation. Submitted

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

VanRaden, P.M. 2007. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91*, 4414-4423.