# The Contribution of Linkage and Linkage Disequilibrium Information to the Accuracy of Genomic Selection

*T. Luan*[*], J.A. Woolliams[†*] and T.H.E. Meuwissen[*]

## Introduction

Substantial advances in genotyping technology have been achieved over the past decade. With the availability of genome-wide dense molecular markers, genomic selection (GS) has now become practical. In this approach, genome-wide breeding values (GW-EBV) are predicted through the use of dense markers covering the whole genome (Meuwissen et al. (2001)). It is distinct from traditional breeding value estimation through the combination of phenotypic data and pedigree information. Marker genotypes for thousands of loci across the whole genome allow GS to predict genetic value more precisely than traditional selection method (Meuwissen (2007)).

Applications of GS to dairy cattle have been performed in North American Holstein (Van Raden et al. (2009)), Australian Holstein-Friesian (Hayes et al. (2009)) and New Zealand Holstein-Friesian and Jersey dairy cattle (Harris et al. (2008)). Accuracies of GW-EBV in these experiments were as high as up to 0.82, which made dairy cattle industry adopt this new technology.

The basic principle of GS is that given a sufficiently high marker density, each quantitative trait loci (QTL) is in linkage disequilibrium (LD) with at least one nearby marker, and a high fraction of the genetic variance is expected to be explained by the markers. Recently, Habier et al. (2007) analyzed accuracies of GW-EBVs resulted from genetic relationships captured by markers. They showed that GS implicitly also used linkage analysis (LA) information. In addition, information of identical-in-state markers may provide LD information in the founders of the pedigree, since the markers may be shared through common ancestors earlier than those in the known pedigree.

The objective of this study was to show how much of the accuracy of GW-EBV is due to LA information and how much due to LD that already existed in the founders of the pedigree. The accuracy of GS applied to British Holstein dairy cattle was investigated for the phenotypes of milk production traits. Three methods, one method based on best linear unbiased prediction, one Bayesian method, and a mixture model approach were used in the study. To calculate the contribution of LA information to the accuracy of GW-EBV, a G-matrix was set up using only LA information as described by Fernando and Grossman (1989).

[*] Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås, N-1432, Norway
[†] The Roslin Institute and R(D)SVS, University of Edinburgh, Roslin, Midlothian EH25 9PS, UK

## Material and methods

**Genotypic and phenotypic data**. Three hundred and thirty British Holstein bulls were genotyped with Illumina Infinium BovineSNP50 BeadChip, which included 53,032 single nucleotide polymorphism (SNP) markers. A total of 45,888 SNPs remained after removing non-segregating SNP and those with Hardy Weinberg equilibrium P<0.01%. The requirement that a bull had at least 45 daughters gave 255 bulls selected for the study. The phenotypic data of all 255 bulls are daughter-yield deviations (DYDs) for the traits: kilograms of milk yield, kilograms of milk fat yield and kilograms of milk protein yield. The average number of daughters per bull is 200.

**Training data.** To obtain the training data sets, the phenotypes of a defined number of individuals were masked, i.e. setting the phenotype "unknown". In the study, we randomly selected 51 individuals at a time, without replacement, to produce 5 non-overlapping training data sets, i.e. every phenotype was masked precisely once in the training data sets. The DYDs of the masked individuals in the training data sets were predicted by GS methods and by Fernando and Grossman method using only LA information. The correlation coefficient between the predicted and realized DYDs was calculated and used as a measure of the accuracy of the EBV predictions.

**Statistical model of GS.** Three models were used to estimate the marker effects: best linear unbiased prediction (G-BLUP), Bayesian statistics (BayesB), and MIXTURE. The general statistical model can be expressed as:

$$\mathbf{y} = \mu + \sum_{j=1}^{N_m} \mathbf{X}_j a_j + \mathbf{e}$$

where **y** is a vector of phenotypes (DYDs), $\mu$ is the overall mean, $N_m$ is the number of markers fitted, $a_j$ is the effect of the marker, $\mathbf{X}_j$ is a vector denoting the genotype of the individuals for marker $j$ and **e** is the random residual vector.

GW-EBV of an individual $i$ was calculated by the second term in the expression as:

$$\text{GW-EBV}_i = \sum_{j=1}^{N_m} X_{ij}\hat{a}_j$$

For G-BLUP, BayesB and MIXTURE, different assumptions are made for the variance of $a_j$. G-BLUP makes no prior assumption on the number of markers explaining the genetic variance, and hence $a_j$ is assumed to be $V_s/N_m$. BayesB assumes that the big SNPs have t-distribution as a prior and the small SNPs have a normally distributed prior with small variance. The prior probability for a SNP being big is $\gamma_B$. MIXTURE model assumes that the marker effects come from a mixture of two distributions: one distribution with large variance (accommodating large marker effects) and one with small variance (accommodating small marker effects), the probability of belonging to the big variance distribution is estimated from the data. The variances of $a_j$ equals $\sigma_1^2$ or $\sigma_2^2$, depending on whether the marker effect is small or large, and $\sigma_1^2$ or $\sigma_2^2$ are both estimated. By adding $\mu$ to the GW-EBV$_i$, and assuming **e** is on average 0, the phenotype of a masked bull can be predicted.

**Prediction of EBV with only LA information.** To predict EBV with only LA information, first, multilocus iterative peeling (Meuwissen (2006); Meuwissen and Goddard (2010)) was

used to estimate the probability that the paternal allele was inherited from the sire and dam at every marker position using Fernando and Grossman (1989). Five generations of pedigree were used for the iterative peeling. The probability of maternal inheritance equals one minus the probability of paternal inheritance. Second, the probabilities of paternal inheritance were used to set up a G-matrix at every marker position. Third, the G matrices were averaged across all marker positions and chromosomes, to obtain an overall G-matrix, $\mathbf{G}_{all}$. $\mathbf{G}_{all}$ was inverted and used by ASReml (Gilmour et al. (2006)) to predict EBV of masked and non-masked individuals.

## Results and discussion

It has been observed that the accuracy for GW-EBV prediction with BayesB was affected little by the prior assumption about the number of markers with an effect (Luan et al. (2009)) for these traits. After some preliminary tests, we set the number of effective genes to 2400, 1600 and 3200 for fat yield, milk yield and protein yield, respectively. The corresponding $\gamma_B$ values are presented in Table 1. Table 1 also lists $\sigma_1^2$ and $\sigma_2^2$ values for MIXTURE as the variances of the two distributions underlying the model.

Table 2 presents the accuracy of the GW-EBV prediction by the G-BLUP, BayesB and MIXTURE methods and the accuracy of the EBV prediction by using only LA information. The table shows for the three models to predict GW-EBV, G-BLUP gives overall the highest accuracy. This agrees with the results of GS study for Norwegian Red cattles (Luan et al. (2009)). It is notable that for fat yield, BayesB achieved a high accuracy, which can be explained by the presence of the DGAT1 gene (Gristart et al. (2002)) in the Holstein. It seems that for a trait whose genetic variance can be explained by a small number of genes, BayesB is preferred.

The contribution of LA information to the reliability, i.e. the square of the accuracy of GW-EBV prediction was calculated as the ratio of the reliability of EBV prediction with only LA information using ASReml with variance component model to the reliability of G-BLUP, and is 0.752, 0.906 and 0.808 for fat, milk and protein yield, respectively. Hence the LA information contributes ~80% of the G-BLUP reliability, and an extra ~20% is achieved by LD that already existed in the founders of the pedigree. For the traits under study, the reliability for fat yield seemed to depend least on LA information, which may be due to the large effect of DGAT1.

For the prediction of the LD, and of the reliability of GS, the effective population size is an important parameter, but effective population size differs at different points in the past. Our results show that the very recent population structure allows for ~80% of the reliability of GS, which suggest that the recent effective population size is most relevant for the prediction of the reliability of GS.

**Table 1: Values of $\gamma_B$ for BayesB and $\sigma_1^2$ and $\sigma_2^2$ for MIXTURE**

| Traits | BayesB | MIXTURE | |
|---|---|---|---|
| | $\gamma_B$ | $\sigma_1^2$ | $\sigma_2^2$ |
| Fat yield | 0.052 | $2.8 \times 10^{-4}$ | $6.7 \times 10^{-2}$ |
| Milk yield | 0.035 | $1.5 \times 10^{-2}$ | $8.4 \times 10^{-2}$ |
| Protein yield | 0.070 | $1.8 \times 10^{-4}$ | $5.4 \times 10^{-2}$ |

**Table 2: Accuracy of GW-EBV prediction and EBV prediction with LA information**

| Methods | Fat yield | Milk yield | Protein yield |
|---|---|---|---|
| G-BLUP | 0.621 | 0.395 | 0.474 |
| BayesB | 0.621 | 0.369 | 0.454 |
| MIXTURE | 0.605 | 0.385 | 0.464 |
| LA | 0.538 | 0.376 | 0.426 |

## Conclusion

These results show that the LA information, which is determined by the pedigree information of the bulls, contributes about 80% of the reliability of GS. This means that although GS in principle doesn't require the pedigree data available, it does use the available population structure, since it is actually included in the marker information. Perhaps the situation with more family data, the contribution of LA is even bigger than what we found here. The findings of the study may explain why the prediction equations derived in one breed do not predict accurate GW-EBV when applied to other breeds, because the family structure is not relevant for the other breed. Although 80% of the reliability of GW-EBV comes from LA information, the extra 20% coming from LD information should not be ignored when applying genomics selection.

## References


Fernando, R.L. and Grossman, M. (1989). *Genet. Sel. Evol.*, 21:467–477.

Grisart, B., Coppieters, W., Farnir, F. *et al.* (2002). *Genome.Res.*, 12:222–231.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., and Thompson, R. (2006). ASReml User Guide Release 2.0

Habier, D., Fernando, R.L., and Dekkers, J.C.M. (1988). *Genetics*, 177: 2389–2397.

Harris, B.L., Johnson, D.L., and Spelman, R.J. (2008). *In Proc Interpol Meeting, Niagara Fall, USA*.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., and Goddard, M.E. (2009). *J. Dairy. Sci.*, 92:433–443.

Luan, T., Woolliams, J.A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T.H.E. (2009). *Genetics*, 183: 1119–1126.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). *Genetics* 157: 1819–1829.

Meuwissen, T.H.E. (2006). In *Proc 8th WCGALP*, volume 20, pages 12.

Meuwissen, T.H.E. (2007). *J. Anim. Breed. Genet.*, 124:321–322.

Meuwissen, T.H.E. and Goddard, M.E. (2010). *Genetics*, accepted