J. Dairy Sci. 98:1–13
http://dx.doi.org/10.3168/jds.2015-9360

Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds

B. C. D. Cuyabano,*[1] G. Su,*[1] G. J. M. Rosa,† M. S. Lund,* and D. Gianola†
*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark
†Department of Animal Sciences, University of Wisconsin, Madison 53706

ABSTRACT

This study compared the accuracy of genome-enabled prediction models using individual single nucleotide polymorphisms (SNP) or haplotype blocks as covariates when using either a single breed or a combined population of Nordic Red cattle. The main objective was to compare predictions of breeding values of complex traits using a combined training population with haplotype blocks, with predictions using a single breed as training population and individual SNP as predictors. To compare the prediction reliabilities, bootstrap samples were taken from the test data set. With the bootstrapped samples of prediction reliabilities, we built and graphed confidence ellipses to allow comparisons. Finally, measures of statistical distances were used to calculate the gain in predictive ability. Our analyses are innovative in the context of assessment of predictive models, allowing a better understanding of prediction reliabilities and providing a statistical basis to effectively calibrate whether one prediction scenario is indeed more accurate than another. An ANOVA indicated that use of haplotype blocks produced significant gains mainly when Bayesian mixture models were used but not when Bayesian BLUP was fitted to the data. Furthermore, when haplotype blocks were used to train prediction models in a combined Nordic Red cattle population, we obtained up to a statistically significant 5.5% average gain in prediction accuracy, over predictions using individual SNP and training the model with a single breed.

Key words: bootstrap analysis, haplotype block, multi-breed genomic prediction, Nordic Red cattle

INTRODUCTION

Genome-enabled prediction methods based on markers, such as SNP, have been widely explored in animal breeding since such methods were introduced (Meuwissen et al., 2001). Haplotype blocks (haploblocks) have been extensively studied in human genetics (Curtis et al., 2001; Chapman et al., 2003; Curtis, 2007), and were explored in animal breeding in the early days of genomic prediction. With advances in SNP technology, a trend emerged toward using SNP genotypes in genome-enabled prediction. More recently, there is renewed interest in the use of haploblocks in prediction of livestock traits for various reasons, one of them being use of information across breeds. Studies have suggested that use of haploblocks can lead to a higher prediction accuracy than use of SNP markers (Villumsen et al., 2009; Boichard et al., 2012; Cuyabano et al., 2014).

A basic assumption in genome-enabled selection is that each QTL is in linkage disequilibrium (**LD**) with some surrounding markers, given that marker panels are sufficiently dense. Genome-enabled prediction uses this assumption as a basis for defining models that estimate the effects of markers on a phenotype of interest. A main potential advantage of haploblocks (i.e., a group of nearby SNP) over individual SNP markers is that each haploblock may be in higher LD with the causative mutations than would be any individual SNP.

One important feature of haploblocks over individual SNP for prediction is that alleles within a haploblock may capture more variation. As a simple example to illustrate this, take 2 biallelic loci A and B of a phased genotype. In the standard regression on SNP genotypes, phenotype y is fitted by equation $\hat{y} = \hat{\mu}_{SNP} + \hat{g}_A A_1 + \hat{g}_B B_1$, where A_1 is 1 if locus A has allele 1 and 0 otherwise, and the same applies to B_1; $\hat{\mu}_{SNP}$ is the mean phenotype when alleles A_2 and B_2 are observed; and $\hat{g}$ are the estimated effects of the respective loci. In a regression on haploblock genotypes, phenotype y is fitted as $\hat{y} = \hat{\mu}_{hap} + \hat{g}_1 A_1 B_1 + \hat{g}_2 A_1 B_2 + \hat{g}_3 A_2 B_1$, where $A_i B_j$ is 1 if locus A has allele i and locus B has allele j (i,j = 1,2) observed, and zero otherwise. When dealing with haploblocks built based on LD, the grouping of many SNP into a haploblock can reduce the number of variables needed to perform genomic prediction (Cuyabano et al., 2014). Furthermore, LD-based haploblocks do not have a fixed number of SNP per haploblock. Table 1

Received January 20, 2015.
Accepted June 16, 2015.
[1]Corresponding authors: beatriz.cuyabano@mbg.au.dk and Guosheng.su@mbg.au.dk

Table 1. Example of estimated haplotype effects by regressing on SNP or haplotype block (haploblock) genotypes, where $\hat{g}$ are the estimated SNP/haploblock effects obtained by the models

	Haplotype effect on phenotype y	
Haplotype	Individual SNP	Haploblocks
A_1B_1	$\hat{g}_A + \hat{g}_B$	$\hat{g}_1$
A_1B_2	$\hat{g}_A$	$\hat{g}_2$
A_2B_1	$\hat{g}_B$	$\hat{g}_3$
A_2B_2	0	0

indicates the estimated effects for regressions on both SNP and haploblock genotypes of each haplotype allele on phenotype y. Using the regression on SNP genotypes, the effect of A_1B_1 is the sum of the effects of A_1B_2 and A_2B_1, whereas haplotypes may have effects that are not linear (e.g., interactions). This may influence accuracy of predictions.

Our hypothesis is, therefore, that haploblocks may improve prediction of economically important traits. To test this, genome-enabled predictions obtained using haploblocks were evaluated and compared with those obtained by using a regression on SNP genotypes. Furthermore, use of haploblocks in an across-breed scenario for prediction may increase predictive ability, mainly due to an increase in the variation captured by haploblock alleles. When working with a single breed, we may not observe some haploblock alleles that occur only in another breed. Hence, the use of a combined population allows us to estimate the effects of haploblock alleles that might not be observed in a single breed analysis.

It is known that when different populations have a common origin, combining them for the training of prediction models can improve prediction reliability (Lund et al., 2011). The stronger the genetic ties between the populations, the greater the benefits (Su et al., 2009). With regard to combined data of different breeds, an increase in prediction reliabilities was reported for Danish, Finnish, and Swedish Red cattle populations when training was based on the combined data instead of using individual breeds (Brøndum et al., 2011). Hence, our work also aimed to assess the potential benefit from using a combined population for training models to obtain genomic predictions of breeding values.

This work presents results of predictions for 5 traits in dairy cattle production: fertility, mastitis, and yields of protein, fat, and milk. We compared the predictive ability of models using regressions on SNP or on haploblock genotypes, and using either a single breed or a combined population of Nordic Red cattle (including Danish, Finnish, and Swedish Red) for model training. We were interested in comparing results when using a Bayesian BLUP model or a Bayesian mixture model to infer allelic effects, to verify whether the model would have an influence on the prediction results. We also wished to measure uncertainty associated with results, which is crucial for proper statistical interpretation. Thus, bootstrap sampling was used in the test data set to generate samples of prediction reliabilities. Samples were analyzed with multivariate statistical methods in an innovative comparison of prediction reliabilities.

MATERIALS AND METHODS

Marker and Phenotypic Data

The marker and phenotypic data used was from a sample of 4,403 animals from the Nordic Red cattle population. The original marker data were obtained from a 54K (∼54,000 markers) SNP chip and then imputed to 777K data using Beagle (Browning and Browning, 2009; Brøndum et al., 2012; Su et al., 2012). The imputed data were then edited by removing markers in complete LD with adjacent markers and with a minor allele frequency (MAF) <0.01 (Su et al., 2012). After editing, a final marker data set with 442,267 SNP for the 4,403 animals was obtained. This population comprised 3 sub-breeds: Danish Red (**DR**), Swedish Red (**SRB**), and Finnish Ayrshire (**FAY**). The data set was split into training and test data sets, using as cut-off the birth date of bulls of October 1, 2001, resulting in a training population of 3,423 animals (663 DR, 1,051 SRB, and 1,709 FAY) and a test population of 980 animals (186 DR, 306 SRB, and 488 FAY).

The phenotypic values used to obtain genome-enabled predictions of fertility, mastitis, protein, fat and milk yield were deregressed proofs (**DRP**) of the bulls, derived from the EBV and their effective daughter contributions (Jairath et al., 1998; Schaeffer, 2001; Garrick et al., 2009).

Animal Ethics

The phenotypic data were collected from routine records of dairy cattle farms. Genotyped animals used in this work were progeny-tested bulls, and the semen samples for genotyping were obtained from routine bull semen collection. Therefore, no ethical approval was necessary.

Haplotype Blocks

The haploblocks were built based on LD measured as D' (Gabriel et al., 2002; Cuyabano et al., 2014). The use of LD for defining haploblocks allows the latter to differ in number of SNP per haploblock, instead of ar-