Haplotype Based Genome-Enabled Prediction of Traits Across Nordic Red Cattle Breeds

Beatriz C. D. Cuyabano, Mogens Sandø Lund, Guilherme Jm Rosa, Daniel Gianola, Guosheng Su

Haplotype Based Genome-Enabled Prediction of Traits Across Nordic Red Cattle Breeds

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

ABSTRACT: SNP markers have been widely explored in genome based prediction. This study explored the use of haplotype blocks (haploblocks) to predict five milk production traits (fertility, mastitis, protein, fat and milk yield), using a mix of Nordic Red cattle as reference population for training. Predictions were performed under a Bayesian approach comparing a GBLUP and a mixture model. In general, predictions were more reliable when using haploblocks instead of individual SNPs as predictors. The Danish Red cattle presented the largest benefit in predictive ability from haploblocks, achieving 5.1% higher reliability than with the individual SNP approach in mastitis. This work gives evidence that predictions using haploblocks along with a combined training population of dairy cattle, may improve prediction accuracy of important traits in the individual populations.
Keywords: Genome-enabled prediction; Haploblocks; Across breed

Introduction

Genome-enabled prediction methods based on single markers, in the form of single nucleotide polymorphisms (SNP), have been widely explored in animal breeding since such methods were introduced (Meuwissen et al., 2001). More recent studies indicate that use of haplotype blocks (haploblocks) can lead to higher prediction accuracy than individual SNP markers in animal breeding (Villumsen et al., 2008; Boichard et al., 2012). Haploblocks have been extensively studied in human genetics (Chapman et al., 2003; Curtis et al., 2001), and are now starting to be explored in animal breeding.

One of the main advantages of haploblocks over individual SNP markers is that they may be in higher LD than any single SNP with causative mutations. Haploblocks are more sensitive to individual SNP changes in the genome (Curtis et al., 2001), and thus may better capture the mutations that affect traits of interest.

Another important advantage of haploblocks over individual SNPs, is that the effects of haplotype-alleles within a haploblock can be assumed independent of each other, better capturing their effects. As a simple example to illustrate this, take two bi-allelic loci A and B. In an individual SNP approach, 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 one if loci A has allele 1 and zero otherwise, and the same rule applies to B_1. In a haploblock approach, phenotype *y* is fitted by equation $\hat{y} = \hat{\mu}_{hap} + \hat{g}_1 A_1B_1 + \hat{g}_2 A_1B_2 + \hat{g}_3 A_2B_1$, where A_iB_j is one if loci A has allele *i* and loci B has allele *j* (*i*,*j*=*1*,*2*) is observed, and zero otherwise.

Table 1 indicates the estimated effects of each haplotype-allele in relation to phenotype *y*, for both individual SNP and haploblock approaches. Note that using the individual SNP approach, the effect of A_1B_1 is the sum of the effects of A_1B_2 and A_2B_1. This may cause confounding of effects, and influence accuracy of predictions.

Table 1. Example of estimated haplotype effects using individual SNP or haploblock approaches.

	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

We hypothesized that haploblocks may improve prediction of economically important traits. Hence, genome-enabled predictions using haploblocks are evaluated and compared to those using individual SNPs.

Another aim of this work was to benefit from using a combined population for training prediction models to genomic breeding values. It is already known that combining populations can improve prediction reliability, when populations have a common origin. The stronger the genetic ties between the populations, the bigger the benefits (Su et al., 2009). In particular, increase in prediction reliabilities has been reported for Danish, Finnish and Swedish Red cattle, when training was based on the combined data, instead of using individual breeds (Brøndum et al., 2011).

The use of haploblocks in across breed prediction has a potential for increasing predictive ability, mainly due to more variation of haplotype-alleles, allowing to estimate effects of haplotype-alleles that might not be observed in a single breed analysis.

This work presents the results on predictions for five traits related to milk production (fertility, mastitis, protein, fat and milk yield), using a combined population of Nordic Red cattle (including Danish, Finnish and Swedish Red) for model training. The aim was to compare the predictive ability of models using individual SNPs or haploblock approaches, as well as compare results when using a GBLUP or a Bayesian mixture model to estimate effects.

Materials and Methods

Data set. The data set consisted of genomic and phenotypic data on 4,366 bulls from the Nordic Red cattle population. This population was a combination of Danish Red (DR, 849 animals), Finnish Ayshire (FAY, 2,178 animals) and Swedish Red (SRB, 1,339 animals) cattle. Because not all the animals had phenotypic records for every trait studied, the size of the training and test data sets varied slightly by trait. Training and test data set sizes, for each population, are described in Table 2.

Table 2. Number of animals in the Red cattle training and test data sets.

		Test animals in Red populations			
Trait	**Train**	**DR**[1]	**FAY**[2]	**SRB**[3]	**All**[4]
Fertility	3,416	186	460	273	919
Mastitis	3,372	186	469	281	936
Protein	3,416	186	460	288	934
Fat	3,416	186	460	273	919
Yield	3,416	186	460	273	919

[1] Danish Red
[2] Finnish Ayshire
[3] Swedish Red
[4] All animals in the test data set

The phenotypic records were de-regressed proofs (DRP) derived from estimated breeding values and effective daughter contributions (Jairath et al., 1998).

Haplotyping method. Haploblocks were built based on linkage disequilibrium (LD) measured as D', and based on the algorithm proposed by Gabriel et al. (2002).

The use of LD to define haploblocks allows them to differ in number of SNPs, instead of arbitrarily defining haploblocks with a fixed length. This non-random setting of where in the genome a haploblock begins and ends, reduces the total number of explanatory variables to be included in the prediction models.

Our previous study investigated the effect of haplotypes, built with different D' thresholds, on prediction and suggested that the optimum for prediction is that every two SNPs in a haploblock must satisfy D' > 0.45. Therefore this D' was used to construct haploblocks in this study.

Prediction models. Prediction was performed for fertility, mastitis, protein, fat and milk yield, using both a genomic best linear unbiased prediction (GBLUP) (Meuwissen et al., 2001) and a Bayesian mixture model (Gao et al., 2013). Both models were implemented with a Bayesian approach using the BayZ package (Janss, 2013).

The models were given by the equation $\mathbf{y} = \mathbf{1}\mu + \mathbf{Mg} + \mathbf{Za} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ was the vector of phenotypic values, μ was the general mean, $\mathbf{M}$ was the genotype matrix (SNPs or haploblocks), $\mathbf{g}$ was the vector of additive genetic effects $\mathbf{Z}$ was the incidence matrix linking $\mathbf{a}$ to $\mathbf{y}$, $\mathbf{a}$ was the vector of residual polygenic effects and $\boldsymbol{\varepsilon}$ was a vector of residuals.

The GBLUP model assumed $\mathbf{g} \sim N(\mathbf{0},\mathbf{I}\sigma_g^2)$ while the mixture model assumed $\mathbf{g} \sim \pi_1 N(\mathbf{0},I\sigma_1^2) + \pi_2 N(\mathbf{0},I\sigma_2^2) + \pi_3 N(\mathbf{0},I\sigma_3^2) + \pi_4 N(\mathbf{0},I\sigma_4^2)$, such that $\sigma_1^2 < \sigma_2^2 < \sigma_3^2 < \sigma_4^2$. The mixing proportions were fixed as $\pi_1 = 0.889$, $\pi_2 = 0.1$, $\pi_3 = 0.01$ and $\pi_4 = 0.001$. For both GBLUP and mixture models, $\mathbf{a} \sim N(\mathbf{0},\mathbf{A}\sigma_a^2)$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\mathbf{D}\sigma_\varepsilon^2)$ for priors. All variances σ_g^2, σ_1^2, σ_2^2, σ_3^2, σ_4^2, σ_a^2, σ_ε^2, and the overall mean μ were assumed to be uniformly distributed. Variances and effects were estimated simultaneously.

Analysis of predictions. After prediction was performed, the genomic estimated breeding values (GEBV) of each individual were obtained as $GEBV_i = \Sigma_j m_{ij}\hat{g}_j + \hat{a}_i$. Predictive abilities were then compared according to their relative prediction reliabilities for test individuals, given by $r^2_{GEBV} = Corr^2(DRP,GEBV)/ r^2_{DRP}$, where. r^2_{DRP} is the average reliability of DRP of animals in the test data set (Garrick et al., 2009).

Results and Discussion

Predictions using haploblocks, generally improved reliabilities, except for protein-GBLUP and fat-mixture, when the test data set was the combined population of the three Red breeds, as shown in Table 3. For these two cases, the reliability of prediction using haploblocks decreased 0.4% for protein and 0.8% for fat, compared to predictions using the individual SNPs. Increase in prediction reliabilities was generally low, but mastitis presented some improvement. This trait had a gain of 2.2% from haploblock prediction compared with individual SNP prediction using a mixture model. The improvements for all the other traits were lower than 1%.

Table 3. Prediction reliabilities for all Red cattle.

Trait	Approach	GBLUP	Mixture
Fertility	Individual SNPs	0.276	0.281
	Haploblocks	0.276	0.284
Mastitis	Individual SNPs	0.258	0.255
	Haploblocks	0.262	0.277
Protein	Individual SNPs	0.360	0.362
	Haploblocks	0.356	0.370
Fat	Individual SNPs	0.446	0.473
	Haploblocks	0.452	0.465
Yield	Individual SNPs	0.371	0.378
	Haploblocks	0.372	0.382

When assessing the predictive ability within each breed, it was mainly the DR cattle that benefited from haploblock predictions (results in Table 4), and most notably when using a mixture model (gain of 5.1% in mastitis, 3.1% in protein, 1.3% in fat and 1.4% in yield). Mastitis and protein presented much bigger gain in reliability of predictions from the haploblock approach when using a mixture model, than using a GBLUP model, in which the gains were less than 1%.

For FAY cattle, differences in reliabilities between haploblock and individual SNP approaches were in general small (results in Table 5), and half of the results were worse than using individual SNPs. Although most of the decrease in predictive ability from haploblocks was lower than 1%, for fat using the mixture model, this decrease was of 1.9%. The best improvement of reliability was observed in mastitis-mixture (gain of 0.9%). Hence, when analyzing the results of haploblocks performance in prediction of FAY

cattle, except for mastitis, there was no benefit over the individual SNP approach.

Table 4. Prediction reliabilities for Danish Red.

Trait	Approach	GBLUP	Mixture
Fertility	Individual SNPs	0.143	0.135
	Haploblocks	0.151	0.124
Mastitis	Individual SNPs	0.331	0.305
	Haploblocks	0.334	0.356
Protein	Individual SNPs	0.339	0.343
	Haploblocks	0.346	0.374
Fat	Individual SNPs	0.354	0.398
	Haploblocks	0.364	0.411
Yield	Individual SNPs	0.356	0.371
	Haploblocks	0.365	0.385

Table 5. Prediction reliabilities for Finnish Ayshire.

Trait	Approach	GBLUP	Mixture
Fertility	Individual SNPs	0.307	0.325
	Haploblocks	0.305	0.324
Mastitis	Individual SNPs	0.244	0.247
	Haploblocks	0.248	0.256
Protein	Individual SNPs	0.330	0.332
	Haploblocks	0.323	0.330
Fat	Individual SNPs	0.500	0.512
	Haploblocks	0.503	0.493
Yield	Individual SNPs	0.350	0.349
	Haploblocks	0.350	0.352

Table 6. Prediction reliabilities for Swedish Red.

Trait	Approach	GBLUP	Mixture
Fertility	Individual SNPs	0.340	0.337
	Haploblocks	0.333	0.359
Mastitis	Individual SNPs	0.214	0.221
	Haploblocks	0.220	0.239
Protein	Individual SNPs	0.409	0.403
	Haploblocks	0.405	0.421
Fat	Individual SNPs	0.410	0.446
	Haploblocks	0.416	0.438
Yield	Individual SNPs	0.408	0.419
	Haploblocks	0.408	0.427

Among predictions for SRB (results in Table 6), three out of ten predictions using haploblocks presented worse reliabilities than the individual SNP approach. However, the decrease in prediction reliabilities were lower than 1% in all cases. As observed in DR, the best gains of haploblocks in predictive ability were noticed when using the mixture model (gain of 2.2% in fertility, 1.8% in mastitis and protein and 0.8% in milk yield).

The main gain in reliabilities of prediction across Nordic Red cattle was observed for mastitis. The predictions for this trait benefited from the use of haploblocks as predictor variables, in either a combined or in the individual populations (DR, FAY or SRB) and for both GBLUP and mixture models. The best reliabilities of prediction were observed when using the mixture model.

Conclusion

This study based on the data across Nordic Red populations gives some evidence about potential benefits from using haploblocks as predictors for complex traits in dairy cattle. Even though not all traits studied presented some improvement from use of haploblocks, most of predictions in DR and SRB had higher reliabilities when using haploblocks instead of individual SNPs.

The increases of 5.1% and 3.1% in prediction reliability of mastitis and protein for DR, using haploblocks instead of individual SNPs, using the mixture model, are the most remarkable results of this work.

In summary, the results suggest that a haploblock approach may enhance accuracy of genome-based prediction across breeds.

Literature Cited

Boichard, D., Guillaume, F., Baur, A. et al. (2012). *Anim. Prod. Sci.* **52**:115-120.

Brøndum, R. F., Rius-Vilarrasa, E., Strandén, I. et al. (2011). *J. Dairy Sci.* **94**:4700-4707.

Chapman, J. M., Cooper, J. D., Todd, J. A. et al (2003). *Hum. Hered.* **56**:18-31.

Curtis, D., North, B. V. and Sham, P. C. (2001). *Ann. Hum. Genet.* **8**:49.

Gabriel, S. B., Schaffner, S. F., Nguyen, H. et al. (2002). *Science* **296**:2225-2229.

Gao, H., Su, G. Janss, L. et al. (2013). *J. Dairy Sci.* **96**:4678-4687.

Garrick, D. J., Taylor, J. F. and Fernando, R. L. (2009). *Genet. Sel. Evol.* **41**:55

Jairath, L. Dekkers, J. C. M., Schaeffer, L. R. et al. (1998). *J. Dairy Sci.* **81**:550:562.

Janss, L. (2013). *BayZ Manual*, version **2.04** ed.

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

Su, G., Guo, G., and Lund, M. S. (2009). *Abstracts of the 60th EAAP*, page 296.

Villumsen T. M., Janss, L. and Lund, M. S. (2008). J. *Anim. Breed. Genet.* **126**:3-13.