Accurate Prediction of Genomic Breeding Values in Norwegian Red Cattle using Dense SNP Genotyping

Tu Luan1, John A. Woolliams1,2, Sigbjørn Lien1, Matthew Kent1,2, Morten Svendsen1 and Theo H.E. Meuwissen1
1Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås, N-1432, Norway
2The Roslin Institute (Edinburgh), Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, Midlothian EH25 9PS, UK
3Centre for Integrative Genetics, Norwegian University of Life Sciences, Ås, N-1432, Norway
4Genes Breeding and A.I. Association, P.O. Box 5021, 1432 Ås, Norway

Introduction

- Genomic selection is a newly developed tool for the estimation of breeding values for quantitative traits through the use of dense markers covering the whole genome.
- With this tool, researchers can check the genetics of an animal even before it is born.
- Studies with simulated data have shown good accuracy could be achieved in estimating genome wide breeding values (GW-EBV).
- Here we investigated the accuracy and possible bias of GW-EBV prediction using real data.

Methods

- G-BLUP – assuming that every marker explains an equal proportion of the total genetic variance.
- BayesB – estimating the variance explained by every marker using a prior distribution.
- MIXTURE – assuming that the marker effects come from a mixture of two distributions with large and small variance.

Real Data

- 500 Norwegian Red bulls – 466 sons of 34 sires, progeny tested between 2001 and 2006 (sons), and before 2001 (sires).
- Genotypic data 18,991 SNPs – genotyped for each bull.
- Phenotypic data Daughter-Yield-Deviations (DYDs) – for production traits (milk yield, fat yield and protein yield) and health traits (calving ease and 3 clinical mastitis defined by period of 1st lactation).

Data Mask

- Phenotype mask Setting the phenotype “unknown” for a defined number of bulls to produce training data set.
  - Random masking
  - Cohort masking
- Masking based on bulls’ year of progeny testing.
- Marker mask Randomly remove markers from the complete data set.
  - 25%, 50% and 75% of 18,991 markers were randomly selected and removed.

Accuracy and Bias

- Accuracy – the correlation coefficient between the predicted and realized DYDs.
- Bias – the regression of the realized phenotypes on the predicted phenotypes.
- Accuracy and bias were estimated by cross-validation.
- For random masking, it is the mean of the accuracies and biases for 5 training data sets.
- For cohort masking, it is a combined accuracy and bias estimated for 50 selected individuals.

Results

Accuracy and bias for production traits and health traits

Accuracy for production traits

- G-BLUP in general achieves the highest accuracy.
- The accuracy for milk yield is lower than that for fat yield and for protein yield.

Bias for production traits

- The degree of bias is judged by comparing the regression coefficient with the value 1.0.
- Prediction for milk yield is mostly less biased.

Accuracy for health traits

- cm1, cm2 and cm3 represent clinical mastitis at 3 different periods of 1st lactation. ce represents calving ease.
- Differences between two maskings for cm1 and cm2 are mostly beyond the standard error of the prediction.

Bias for health traits

- For cm1, cm2 and cm3, prediction with higher accuracy has lower bias.
- GW-EBV predictions for calving ease have highest accuracy, and in general lowest bias.

Accuracy for full data and subsets with different size of markers

Conclusion

- The accuracies of the GW-EBV prediction were found to vary widely between 0.12 and 0.62.
- G-BLUP gave overall the highest accuracy.
- It was observed a strong relationship between the accuracy of the prediction and the heritability of the trait.
- GW-EBV prediction for production traits with high heritability achieved higher accuracy and also lower bias than health traits with low heritability.
- To achieve a similar accuracy for the health traits probably more records will be needed.