Genomic Prediction Using High-Density SNP Markers in Nordic Holstein and Red

Guosheng Su[1], Rasmus F. Brøndum[1], Peipei Ma[1],
Bernt Guldbrandtsen[1], Gert P. Aamand[2], Mogens S. Lund[1]
[1]Department of Molecular Biology and Genetics, Aarhus University, Denmark
[2]Nordic Cattle Genetic Evaluation, DK-8200 Aarhus N, Denmark

Abstract

This study investigated genomic prediction using medium density and high density markers, based on data from Nordic Holstein and Red (RDC). The Holstein data comprised 4,539 progeny-tested bulls, and the RDC data 4,403 bulls. The data were divided into reference data and test data using 2001-10-01 as a cut-off date (birth date of the bulls). This resulted in about 25% genotyped bulls in Nordic Holstein test data, and 20% in RDC test data. For each breed, three datasets of markers were used for predicting breeding values: 1) 50k dataset with some missing markers, 2) 50k dataset where missing markers were imputed, and 3) imputed HD dataset which was created by imputing the 50k data to HD data based on 557 bulls genotyped using 770k chip in Holstein, and 706 bulls in RDC. Based on the three marker datasets, direct genomic breeding values (DGV) for protein, fertility and udder health were predicted using a GBLUP model and a Bayesian mixture model with two normal distributions. Reliability of DGV was measured as squared correlations between de-regressed proofs (DRP) and DGV and then corrected for reliability of DRP, and unbiasedness was assessed by regression of DRP on DGV, based on the bulls in the test datasets. Averaged over the three traits, reliability of DGV based on the HD markers was 0.5% higher than that based on the 50k data in Holstein, and 1.0% higher in RDC. In addition, the HD markers led to an improvement on unbiasedness of DGV. The Bayesian mixture model led to 0.5% higher reliability than the GBLUP model in Holstein, but not in RDC. Compared with the raw 50k data, the imputed 50k data improved genomic prediction for protein in RDC.

Keywords: genomic prediction, high density markers, medium density markers

1. Introduction

One of the important factors affecting accuracy of genomic prediction is marker density. Higher marker density means that on average the markers will be in stronger linkage disequilibrium (LD) with genes affecting the trait of interest, which consequently should lead to better genomic predictions.

Currently a medium density SNP chip with 50k markers is in wide use for genomic prediction in dairy cattle. In 2010, a high density (HD) SNP chip with 770k markers was released. It is expected that the HD markers will lead to more accurate genomic predictions than the 50k markers do. However, simulation studies show that the advantage of HD markers in genomic prediction is large when few genes affect the trait (Meuwissen and Goddard, 2010), but very small in the case of a large number of genes affecting the trait (VanRaden *et al.*, 2011).

Marker-QTL associations differ among populations. The differences are dependent on the genetic distances between populations. Thus the advantage in genomic prediction of changing to HD markers should be more profound for genomic prediction across populations than within population.

The objective of this study is to compare genomic predictions using imputed HD markers and current 50k markers, based on the data from the Nordic Holstein and Red Dairy Cattle (RDC) populations.

2. Material and Methods

2.1 Data

The data used in this analysis were genotypes and de-regressed proofs (DRP) from Nordic Holstein and RDC populations. DRP were derived from genetic evaluations in 2010-11. The traits under analysis were protein, fertility and udder health. The Holstein data comprised 4,539 progeny-tested bulls, and the RDC data 4,403 bulls. The bulls were genotyped using the Illumina Bovine SNP50 BeadChip. Among the RDC bulls, 706 bulls were re-genotyped using 770k chip. For Holstein, 557 bulls in the EuroGenomics project (Lund *et al.*, 2009) were re-genotyped using the HD chip.

The 50k genotypes were imputed to the HD genotypes using Beagle package (Browning and Browning, 2009), based on the marker data of the HD genotyped bulls. The markers in the 50k chip but not included in the HD chip were excluded in the imputation process. After imputation, the markers in complete linkage with the previous markers were removed. In order to investigate the effect of inferring missing marker on genomic prediction, the missing markers in the 50k data were also imputed using Beagle package.

For each breed, three datasets of markers were used for predicting breeding values: 1) raw 50k data with some missing markers, 2) imputed 50k data where missing markers in the 50 data were imputed ($50k_{imp}$), and 3) imputed HD data (HD). In RDC, markers of all 30 chromosomes were used. But in Holstein, the X chromosome was excluded, because this chromosome was not as a part of exchanges in the EuroGenomics project. The number of markers used in genomic prediction is 46,847 in the 50k dataset and 528,595 in the HD dataset for RDC, and 43,415 in the 50k dataset and 492,057 in the HD dataset for Holstein.

2.2 Statistical models

Direct genomic breeding values were predicted using two models. One is GBLUP model, the other is a Bayesian mixture model.

GBLUP: The GBLUP model is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e}$$

where **y** is the vector of DRP, **g** is the vector of DGV, and **e** is the vector of residuals.

It is assumed that $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma^2_g)$, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D}\sigma^2_e)$, where **G** is a genomic relationship matrix, σ^2_g is genomic additive genetic variance, **D** is a diagonal matrix and σ^2_e is residual variance. **G** is defined as $\mathbf{G} = \mathbf{MM}' / \sum 2 p_i q_i$ where elements in column i of **M** are $0 - 2p_i$, $1 - 2p_i$ and $2 - 2p_i$ for genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. **D** has diagonal element $d_{ii} = (1-r^2_{DRP})/r^2_{DRP}$ which is applied to account for heterogeneous residual variances due to different reliabilities of DRP (r^2_{DRP}).

Bayesian mixture: The Bayesian mixture model is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Mq} + \mathbf{e}$$

where **q** is the vector of SNP genotype effects, **M** is defined as the above.

The model assumes that a small proportion (π) of SNP have large effect and the rest have small effect. This is achieved by assuming that prior distribution of q_i is either a normal distribution with a large variance (σ^2_{v1}) or a normal distribution with small variance (σ^2_{v0}), i.e.,

$$q_i \sim (1-\pi)\, N(0, \sigma^2_{v0}) + \pi\, N(0, \sigma^2_{v1})$$

In the present study, π was set to be 0.05, 0.10, 0.20 or 0.50 when using the 50k markers, and 0.005, 0.01, 0.02, or 0.05 when using the HD markers. These settings made such that the expected number of markers to be in the distribution with large variance of the mixture is almost the same when using the 50k markers and the HD markers.

2. 3 Validation

The error rate of imputation from the 50k to the HD markers was assessed by a validation where the HD genotyped bulls were divided into reference and test data. For RDC, the test data contained 150 bulls, and for Holstein, the test

data consisted of 100 bulls. In the test data, the HD markers not in the 50k map were excluded, and then imputed. The error rate was calculated as the proportion of number of false alleles to total number of imputed alleles.

In the validation of genomic prediction, the whole dataset in each breed was divided into reference (training) data and test data by the cut-off date 2001-10-01. Genomic predictions using different marker datasets and using different models was evaluated by comparing DGV and DRP for the animals in the test data. Reliability of DGV was measured as squared correlation between DGV and DRP, and then divided by reliability of DRP. Unbiasedness of genomic prediction was assessed by regression of DRP on DGV.

3. Results

3.1 Imputation error rate

As shown in Table 1, allele error rate of imputation was 0.77% in Nordic Holstein population, and 0.96% for Nordic RDC population. In addition, there was a variation in error rates among the three RDC populations: Danish Red had a higher error rate (1.75%), and Finnish Ayrshire and Swedish Red had lower error rates (0.54% and 0.59%, respectively). The results indicate accurate imputation from the 50k to the HD markers.

3.2 Genomic prediction in Nordic Holstein

Reliabilities of genomic prediction for Holsteins based on the 50k and the HD markers using two alternative models are shown in Table 2. The HD markers led to an increase of reliability of DGV for protein and fertility, but not for udder health. On average, reliability of DGV based on the HD markers was 0.5% higher than that based on the 50k markers. It was observed that the Bayesian mixture model was superior to the GBLUP model, regardless of which marker dataset was used. On average, the increase of reliability using the Bayesian mixture model was 0.5%. On the other hand, imputation of missing markers in the 50k data did not yield any improvement of reliability of DGV.

A necessary condition of unbiased genomic prediction is that the regression coefficient of DRP on genomic prediction is not far from one. As shown in Table 3, the HD markers led to less biased DGV for protein and fertility but not for udder health. The Bayesian model did not reduce bias of DGV. Imputing missing markers in the 50k data slightly increased bias compared to the raw 50k data.

3.3 Genomic prediction in Nordic RDC

The influences of models and marker datasets on reliability of DGV in Nordic RDC (Table 4) are somewhat different from those in Nordic Holstein. Imputing missing markers in the 50k data improved reliability of DGV for protein, but not for the other two traits. The Bayesian mixture model gave very similar reliability as GBLUP, based on the 50k markers, and was slightly better than GBLUP based on the HD markers. Applying the GBLUP model, reliability of DGV using the HD markers was on average 1.0% higher than using the raw 50k markers, and 0.7% higher than using the imputed 50k markers. When applying the Bayesian mixture model, the increase of reliability using the HD markers was 1.20% and 0.80%, respectively, compared with the raw 50k and the imputed 50k markers.

The regression coefficients of DRP on DGV were closer to one when DGV were predicted based on the HD makers, indicating a reduction of bias using HD markers. Similar to the Holstein population, the Bayesian mixture model did not reduce bias of DGV in RDC, regardless of the marker dataset used. But in contrast to Holstein, imputing missing markers in the 50k data reduced bias of DGV.

4. Discussion

This study investigated the advantage of using HD markers for genomic prediction. Based on the present data and models, reliability of DGV based on the HD markers was, on average, 0.5% higher than that based on the 50k data in Holstein, and 1.0% in RDC. In addition, the HD markers led to a reduction of bias in genomic predictions. The results are consistent with simulation studies assuming a large number of

genes affecting the trait. VanRaden *et al.* (VanRaden *et al.*, 2011) reported that increasing number of markers from 50k to 500k yielded a gain of 1.6% in their simulation study. Harris et al. (Harris and Johnson, 2010) reported very little gain when the number of markers increased from 20k to 1000k in a simulation study.

The Nordic RDC in this study including the Finnish Ayrshire, Swedish Red and Danish Red populations. The gain of genomic prediction using the HD markers in RDC was larger than that in Holstein. This supports that HD markers give more benefit for genomic prediction across populations than within population, because the LD between genes with adjacent markers is not well preserved across populations in 50k markers but well in HD markers (Villa-Angulo *et al.*, 2009).

The number of markers in the HD dataset is more than 10 times greater than the 50k dataset, thus there must be a much stronger LD between markers and genes affecting the trait of interest. Therefore it is expected that the HD markers will lead to much better genomic prediction. However, the current study shows that the gain from the HD markers is small. The following could be the possible reasons.

Firstly, the advantage of increasing LD by HD markers may be counteracted by increasing the number of unknown parameters to be estimated. It may be necessary to reduce number of markers by deleting redundant markers which are non-informative for population genome structure. Secondly, the models in this study might not be optimal. More sophisticated variable selection models are required. Thirdly, the HD markers are not real markers genotyped using HD chip, but imputed ones. The actual imputation error rate may be higher than that indicated in the validation analysis, because the validation was based on real HD genotyped animals among which the relationship could be stronger than that the relationship between HD genotyped animals and 50k genotyped animals.

In conclusion, the gain of genomic prediction using HD markers is small, based on current data and models. Further studies are needed before HD markers can be used for practical genomic prediction.

Acknowledgments

We thank Danish Cattle Federation, Faba Co-op, Swedish Dairy Association and Nordic Cattle Genetic Evaluation for providing data. This work was performed in the project “Genomic Selection – from function to efficient utilization in cattle breeding (grant no. 3412-08-02253)”, funded under Green Development and Demonstration Programme by the Danish Directorate for Food, Fisheries and Agri Business, the Milk Levy Fund, VikingGenetics, Nordic Cattle Genetic Evaluation, and Aarhus University.

References

Browning, B.L. & Browning, S.R. 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am. J. Hum. Genet. 84*, 210-223.

Harris, B.L. & Johnson, D.L. 2010. The impact of high density SNP chips on genomic evaluation in dairy cattle. *Interbull Bulletin 42*, 40-43.

Lund, M.S., Sahana, G., de Koning, D.J. , Su, G. & Carlborg, Ö. 2009. Comparison of analyses of the QTLMAS XII common dataset. I: *Genomic selection. BMC Proceedings 3*, s. 1.

Meuwissen, T. & Goddard, M. 2010. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. *Genetics 185*, 623-U338.

VanRaden, P.M., O'Connell, J.R., Wiggans, G.R. & Weigel, K.A. 2011. Genomic evaluations with many more genotypes. *Genetics Selection Evolution 43:10.*

Villa-Angulo, R., Matukumalli, L.K., Gill, C.A., Choi, J., Van Tassell, C.P. & Grefenstette, J.J. 2009. High-resolution haplotype block structure in the cattle genome. *BMC Genet. 10.*

Table 1. Allele error rate of imputation in Holstein and RDC.

Breed	N_ref	N_test	Error rate %
Holstein	457	100	0.77
RDC	556	150	0.96

Table 2. Reliability of DGV for Holstein using GBLUP and Bayesian mixture based on 50k and HD markers.

Trait	N	GBLUP			Bayesian mixture		
		50k	$50k_{imp}$	HD	50k (π=0.2)	$50k_{imp}$ (π=0.2)	HD (π =0.02)
Protein	1395	0.425	0.426	0.429	0.435	0.434	0.440
Fertility	1378	0.404	0.403	0.413	0.406	0.406	0.416
Udder health	1461	0.370	0.372	0.370	0.375	0.376	0.376
Average	1411	0.400	0.400	0.404	0.405	0.405	0.410

Table 3. Regression of DRP on DGV for Holstein using GBLUP and Bayesian mixture based on 50k and HD markers.

Trait	N	GBLUP			Bayesian mixture		
		50k	$50k_{imp}$	HD	50k (π=0.2)	$50k_{imp}$ (π=0.2)	HD (π =0.02)
Protein	1395	0.853	0.847	0.863	0.855	0.845	0.862
Fertility	1378	0.972	0.963	0.994	0.968	0.958	0.996
Udder health	1461	0.952	0.933	0.946	0.948	0.927	0.946
Average	1411	0.926	0.914	0.934	0.924	0.910	0.935

Table 4. Reliability of DGV for RDC using GBLUP and Bayesian mixture based on 50k and HD markers.

Trait	N	GBLUP			Bayesian mixture		
		50k	$50k_{imp}$	HD	50k (π=0.2)	$50k_{imp}$ (π=0.2)	HD (π =0.02)
Protein	923	0.346	0.358	0.358	0.346	0.357	0.359
Fertility	940	0.297	0.293	0.304	0.299	0.296	0.307
Udder health	978	0.244	0.246	0.257	0.243	0.248	0.259
Average	947	0.296	0.299	0.306	0.296	0.300	0.308

Table 5. Regression of DRP on DGV for RDC using GBLUP and Bayesian mixture based on 50k and HD markers.

Trait	N	GBLUP			Bayesian mixture		
		50k	$50k_{imp}$	HD	50k (π=0.2)	$50k_{imp}$ (π=0.2)	HD (π =0.02)
Protein	923	0.849	0.875	0.877	0.835	0.864	0.877
Fertility	940	0.934	0.939	0.980	0.933	0.940	0.980
Udder health	978	0.851	0.854	0.872	0.839	0.846	0.870
Average	947	0.878	0.889	0.910	0.869	0.883	0.909