Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations

G. Su,^{*1} R. F. Brøndum,^{*} P. Ma,^{*†} B. Guldbrandtsen,^{*} G. P. Aamand,[‡] and M. S. Lund^{*}

^{*}Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark

[†]College of Animal Science and Technology, China Agricultural University, 100193 Beijing, China

[‡]Nordic Cattle Genetic Evaluation, DK-8200 Aarhus, Denmark

ABSTRACT

This study investigated genomic prediction using medium-density (~54,000; 54K) and high-density marker panels (~777,000; 777K), based on data from Nordic Holstein and Red Dairy Cattle (RDC). The Holstein data comprised 4,539 progeny-tested bulls, and the RDC data 4,403 progeny-tested bulls. The data were divided into reference data and test data using October 1, 2001, as a cut-off date (birth date of the bulls). This resulted in about 25% genotyped bulls in the Holstein test data and 20% in the RDC test data. For each breed, 3 data sets of markers were used to predict breeding values: (1) 54K data set with missing genotypes, (2) 54K data set where missing genotypes were imputed, and (3) imputed high-density (HD) marker data set created by imputing the 54K data to the HD data based on 557 bulls genotyped using a 777K single nucleotide polymorphism chip in Holstein, and 706 bulls in RDC. Based on the 3 marker data sets, direct genomic breeding values (DGV) for protein, fertility, and udder health were predicted using a genomic BLUP model (GBLUP) and a Bayesian mixture model with 2 normal distributions. Reliability of DGV was measured as squared correlations between deregressed proofs (DRP) and DGV corrected for reliability of DRP. Unbiasedness was assessed by regression of DRP on DGV, based on the bulls in the test data sets. Averaged over the 3 traits, reliability of DGV based on the HD markers was 0.5% higher than that based on the 54K data in Holstein, and 1.0% higher than that in RDC. In addition, the HD markers led to an improvement of unbiasedness of DGV. The Bayesian mixture model led to 0.5% higher reliability than the GBLUP model in Holstein, but not in RDC. Imputing missing genotypes in the 54K marker data did not improve genomic predictions for most of the traits.

Key words: genomic prediction, high-density marker panel, imputation

INTRODUCTION

One of the important factors affecting accuracy of genomic prediction is marker density (Solberg et al., 2008; Habier et al., 2009; Meuwissen, 2009; Weigel et al., 2009). Higher marker density means that, on average, the markers are in stronger linkage disequilibrium (LD) with genes affecting the trait of interest, which should lead to better genomic predictions.

Currently, a medium-density SNP chip with ~54,000 markers (**54K**; Matukumalli et al., 2009) is widely used for genomic prediction in dairy cattle (Su et al., 2010; VanRaden and Sullivan, 2010; Lund et al., 2011). In 2010, a high-density (**HD**) SNP chip with ~777,000 markers (**777K**) was released (Matukumalli et al., 2011). It is expected that using the HD markers will lead to more accurate genomic predictions than using the 54K chip. 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 depend on the genetic distances between populations (Gautier et al., 2007; de Roos et al., 2008, 2009). The more closely related populations are, the more LD patterns are expected to be preserved among the populations. It has been reported that between *Bos taurus* cattle breeds, the LD phase is persistent only for marker pairs less than 10 kb apart (Gautier et al., 2007; de Roos et al., 2008). For the cattle genome, this requires a density of at least 300,000 markers. Thus, the benefit of changing from 54K to HD markers should be more profound for genomic prediction across populations than within populations. In the Nordic dairy cattle joint genetic evaluation, the Red Dairy Cattle (**RDC**) population consists of Finnish Ayrshire, Swedish Red,

Received January 25, 2012.

Accepted April 17, 2012.

¹Corresponding author: guosheng.su@agrsci.dk

and Danish Red. The Holstein population is mainly Danish Holstein. Therefore, the RDC population can be considered as a mixture of 3 populations, whereas the Holstein population can be taken as a single population. This leads to a hypothesis that the benefit for genomic prediction using HD markers rather than 54K markers would be larger in the RDC population than in the Holstein population.

The BLUP model (to estimate either SNP effects or individual additive genetic effects) is a popular approach in practical genomic evaluations using 54K markers (VanRaden et al., 2009; Harris and Johnson, 2010a; Liu et al., 2011; Su et al., 2012), because it is simple, has relatively low computational requirements, and performs as well as variable selection models for most traits (Hayes et al., 2009a; VanRaden et al., 2009). Using HD markers, the number of unknowns in a prediction model increases dramatically. It is expected that variable selection models will predict genomic breeding values better than linear BLUP models because they can better attribute genetic variance to SNP in close LD with the QTL.

The objective of this study was to compare genomic predictions using either imputed HD markers or current 54K markers, applying either a linear BLUP model with genomic relationship matrix (genomic BLUP, **GBLUP**) or a Bayesian mixture model, based on the data from Nordic Holstein and RDC populations.

MATERIALS AND METHODS

Data

The data used in this analysis were genotypes and deregressed proofs (**DRP**) from Nordic Holstein and RDC populations. The DRP were derived from genetic evaluations in November 2010. The traits under analysis were protein yield, fertility, and udder health, which were the economically most important traits in the Nordic total merit index, and varied widely in heritability (from 0.04 for fertility and udder health to 0.39 for protein yield). The Holstein data comprised 4,539 progeny-tested bulls (mainly Danish Holstein), and the RDC data comprised 4,403 bulls (49.5% Finnish Ayrshire, 30.4% Swedish Red, 19.3% Danish Red, and 0.8% imported Red).

The bulls were genotyped using the Illumina Bovine SNP50 BeadChip (Illumina Inc., San Diego, CA). Among the RDC bulls, 706 bulls (about one-third for each of the 3 RDC populations) were re-genotyped using the Illumina BovineHD BeadChip (777K). For Holstein, 557 bulls in the EuroGenomics project (Lund et al., 2011) were re-genotyped using the HD chip. The 54K genotypes were imputed to the HD genotypes using the

Beagle package (Browning and Browning, 2009), based on the marker data of the HD genotyped bulls. Because the aim of this study was to compare the 54K and HD markers for genomic predictions, the imputation was based on the HD map, and those markers on the 54K chip but not on the HD chip were excluded in the imputation process. To investigate the effect of inferring missing genotypes on genomic predictions, the missing genotypes in the 54K data (due to applying different versions of the Illumina 54K chip, and genotypes failing or being of poor quality) were also imputed using the Beagle package. All imputed genotypes were accepted. Thus, there were no missing genotypes in the imputed 54K and HD data. The unimputed 54K data and the imputed 54K data were edited with criteria of minor allele frequency (**MAF**) 0.01 and locus average GenCall score 0.60. The imputed HD data were edited by deleting the markers that were in complete LD with the adjacent markers and the markers with $MAF < 0.01$. To delete the markers in complete LD with the adjacent markers, LD between a marker and the next marker was inspected, starting from the first marker on each chromosome. If a marker (SNP_i) and the next marker (SNP_{i+1}) was in complete linkage, SNP_{i+1} was deleted, and then SNP_i was compared with SNP_{i+2} ; otherwise SNP_{i+1} was compared with SNP_{i+2} . After the procedure was complete, the LD (r^2) of any pair of adjacent markers was < 1 .

For each breed, 3 marker data sets were used to predict breeding values: (1) unimputed 54K data, where missing marker genotypes (3.9% in Holstein and 4.4% in RDC) were replaced with population expectation calculated from allele frequencies at the corresponding locus; (2) imputed 54K data, where missing genotypes in the 54K data were imputed; and (3) imputed HD data. In RDC, markers on all 30 chromosomes were used. In Holstein, the X chromosome was excluded, because this chromosome was not exchanged as part of the EuroGenomics project. Because of small differences in allele frequencies between original and imputed 54K data sets, the numbers of markers in the original and imputed 54K data sets were not the same after deleting markers with minimal $MAF < 0.01$ (Table 1).

Statistical Model

Direct genomic breeding values (**DGV**) were predicted using 2 models. One was a GBLUP model and the other was a Bayesian mixture model.

GBLUP. The GBLUP model (VanRaden, 2008; Hayes et al., 2009b) is

$$y = 1\mu + Zg + e,$$

Table 1. Number of SNP markers before editing (n_{raw}) and after editing (n_{ed}), and average pair-wise linkage disequilibrium (LD) between adjacent markers

SNP panel ¹	Breed	n_{raw} ²	n_{ed} ²	LD ³
54K	Holstein	46,973	43,413/43,922 _(imp)	0.209
	Red Dairy Cattle	49,657	45,168/46,847 _(imp)	0.180
777K	Holstein	648,219	492,057	0.557
	Red Dairy Cattle	673,295	528,595	0.533

¹Medium-density (~54,000 markers; 54K) and high-density (~777,000 markers; 777K) SNP panels.

²Number of markers including X chromosome in Red Dairy Cattle, excluding X chromosome in Holstein. Because of small differences in allele frequencies between original and imputed (imp) 54K data sets, the numbers of markers in original and imputed 54K data sets were not the same after editing.

³Measured as r^2 , calculated based on markers in autosomes, using the SNP marker data before editing.

where $\mathbf{y}$ is the vector of DRP, μ is the overall mean, $\mathbf{1}$ is a vector of 1s, $\mathbf{g}$ is the vector of DGV, $\mathbf{Z}$ is the incidence matrix for $\mathbf{g}$, and $\mathbf{e}$ is the vector of residuals.

It was assumed that $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D}\sigma_e^2)$, where $\mathbf{G}$ is a genomic relationship matrix, σ_g^2 is the genomic additive genetic variance, $\mathbf{D}$ is a diagonal matrix, and σ_e^2 is the residual variance. Matrix $\mathbf{G}$ is defined as $\mathbf{G} = \mathbf{M}\mathbf{M}' / \sum 2p_i q_i$, where elements in column i of $\mathbf{M}$ are $0 - 2p_i$, $1 - 2p_i$, and $2 - 2p_i$ for genotypes A_1A_1 , A_1A_2 , and A_2A_2 , respectively, q_i is the allele frequency of A_1 , and p_i is the allele frequency of A_2 . In theory, base population allele frequencies should be used to construct a $\mathbf{G}$ matrix (Gengler et al., 2007; VanRaden, 2008). However, many studies have shown that allele frequencies observed from current marker data perform as well as estimated base population allele frequencies with regard to accuracy of predicted genomic breeding value (Aguilar et al., 2010; Forni et al., 2011). In this study, allele frequencies observed from the current marker data were used to construct the $\mathbf{G}$ matrix. When using the unimputed 54K data, the missing marker genotype was replaced with the population expectation at the corresponding locus; that is, missing genotypes at locus $j = 0(1 - p_j)^2 + 1[2p_j(1 - p_j)] + 2p_j^2 = 2p_j$, which was equivalent to using zero to replace the elements for missing genotypes in the $\mathbf{M}$ matrix ($2p_j - 2p_j = 0$). In other words, it was equivalent to assume that missing genotypes had null effect. Matrix $\mathbf{D}$ has a diagonal element $d_{ii} = (1 - r_{\text{DRP}}^2) / r_{\text{DRP}}^2$ to account for heterogeneous residual variances due to different reliabilities of DRP (r_{DRP}^2). Variances (σ_g^2 and σ_e^2) used for predictions were those estimated from reference data and the corresponding marker data.

Bayesian Mixture. The Bayesian mixture model (Meuwissen, 2009) is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{M}\mathbf{q} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of DRP, $\mathbf{q}$ is the vector of SNP genotype effects (q_i), and $\mathbf{M}$ is as defined above. The model assumes that a small proportion (π) of SNP has large effects, and the remainder has small effects. This is achieved by assuming that the prior distribution of q_i is either a normal distribution with a large variance (σ_{v1}^2) or a normal distribution with small variance (σ_{v0}^2); that is, $q_i \sim (1 - \pi)N(\mathbf{0}, \sigma_{v0}^2) + \pi N(\mathbf{0}, \sigma_{v1}^2)$.

In the present study, π was set to be 0.05, 0.10, 0.20, or 0.50 when using the 54K markers, and 0.005, 0.01, 0.02, or 0.05 when using the HD markers. These settings were chosen 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 54K markers and the HD markers. The Gibbs sampling algorithm was applied to the Bayesian mixture model. The Gibbs sampler was run as a single chain with a length of 50,000 samples. The first 20,000 samples were discarded as burn-in, and every 10th sample of the remaining 30,000 was saved to calculate posterior statistics. In general, the largest π led to slightly lower prediction accuracy than the other 3 priors in Holstein, and the smallest and the largest π yielded slightly lower prediction accuracy than the other 2 priors in RDC, regardless of 54K or HD data. In the context, the presented results were those from the scenario of $\pi = 0.20$ when using the 54K markers and of $\pi = 0.02$ when using HD markers, which were generally appropriate for the traits in the current study.

Validation

The error rate of imputation from the 54K to the HD markers was assessed by a validation in which 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. The bulls in the test data were randomly chosen from those

Table 2. Heritability (h^2) of the traits, number of bulls (n), and reliability of deregressed proofs (r_{DRP}^2) in reference and test data sets

Breed	Trait	h^2	Reference		Test	
			n	r_{DRP}^2	n	r_{DRP}^2
Holstein	Protein	0.39	3,003	0.940	1,395	0.924
	Fertility	0.04	3,037	0.682	1,378	0.607
	Udder health	0.04	3,005	0.823	1,461	0.749
Red Dairy Cattle	Protein	0.39	3,421	0.947	924	0.917
	Fertility	0.04	3,377	0.786	941	0.671
	Udder health	0.04	3,421	0.905	979	0.797

HD genotyped bulls that did not have HD genotyped sons. In the test data, the HD markers not in the 54K map were deleted, and then imputed. The error rate was calculated as the number of wrongly imputed alleles in proportion to the total number of imputed alleles.

In the validation of genomic predictions, the whole data set in each breed was divided into reference (training) data and test data by the cut-off date (birth date of bulls) on October 1, 2001. The number of bulls in the reference and test data and the average reliability of DRP for each trait are shown in Table 2. The numbers of bulls were somewhat different among the traits. The main reason was that some bulls did not have EBV for one or more traits due to the restriction that the published EBV (from which DRP were derived) for protein should have a reliability of at least 0.60, and for fertility and udder health of at least 0.35.

Genomic predictions using different marker data sets and different models were evaluated by comparing DGV and DRP for animals in the test data. Reliability of DGV was measured as squared correlation between DGV and DRP divided by the reliability of DRP (Lund et al., 2011; Su et al., 2012). Unbiasedness of genomic prediction was assessed by regression of DRP on DGV. Given unbiased predictions, it is expected that the covariance

$$\text{Cov}(DGV, DRP) = \text{Cov}(DGV, DGV + \varepsilon + e) = \sigma_{DGV}^2,$$

where ε is the prediction error of DGV and e is the residual of DRP; thus, the regression coefficient

$$b_{DRP/DGV} = \text{Cov}(DGV, DRP) / \sigma_{DGV}^2 = 1.$$

RESULTS

LD Between Markers and Imputation Error Rate

Based on the SNP marker data before editing, the ratio of the number of markers in the HD marker data to the number in the 54K marker data was about 13.5:1

(Table 1). Correspondingly, average pair-wise distance between adjacent markers was about 4.5 kb in the HD data and 60 kb in the 54K data. This indicates that the density of the HD is higher than the requirement (distance of marker pairs <10 kb) for persistent LD phase between *Bos taurus* breeds (Gautier et al., 2007; de Roos et al., 2008). Average pair-wise LD (r^2) between adjacent markers in the HD marker data was 2.7 times as high as in 54K data for Holstein and 3.0 times for RDC. Linkage disequilibrium was higher for Holstein compared with RDC, regardless of marker data sets. After marker data editing, the ratio of the number of markers in the HD marker data to the number in the 54K marker data was decreased to 11.3:1, because many markers in complete LD with other markers in HD marker data were deleted.

As shown in Table 3, the allele error rate of imputation from the 54K to the HD markers was 0.77% for Holstein, and 0.96% for RDC. In addition, we observed variation in error rates among the 3 RDC populations: Danish Red had a higher error rate (1.75%) than Finnish Ayrshire (0.54%) and Swedish Red (0.59%), although the number of reference bulls was almost the same in each of the 3 RDC populations. The results indicated that imputation from the 54K to the HD markers was quite accurate.

Estimates of Additive Genetic Variances and SNP-Effect Variances

Table 4 presents the estimated additive genetic variances using the GBLUP model and SNP-effect vari-

Table 3. Number of bulls in the imputation reference (n_{ref}) and test data (n_{test}) and allele error rate of imputation from 54K (~54,000 markers) to 777K (~777,000 markers) data

Breed	n_{ref}	n_{test}	Error rate (%)
Holstein	457	100	0.77
Red Dairy Cattle	556	150	0.96

Table 4. Estimates of additive genetic variances (σ_g^2) from the genomic BLUP (GBLUP) model and SNP variances (σ_{v0}^2 and σ_{v1}^2 ; $\times 10,000$) from the Bayesian mixture model¹

Breed	Trait	GBLUP			Bayesian mixture					
		54K	54K _{imp}	777K	54K	54K _{imp}	777K	54K	54K _{imp}	777K
		σ_g^2	σ_g^2	σ_g^2	σ_{v0}^2	σ_{v1}^2	σ_{v0}^2	σ_{v1}^2	σ_{v0}^2	σ_{v1}^2
Holstein	Protein	129.9	129.4	131.0	2.634	140.4	1.977	138.0	0.159	123.7
	Fertility	142.2	138.8	140.8	5.768	143.8	3.607	143.5	0.252	129.2
	Udder	93.2	93.2	93.4	1.505	103.2	0.962	102.8	0.109	88.4
Red Dairy Cattle	Protein	99.7	95.9	97.9	3.600	94.5	3.181	85.8	0.149	81.8
	Fertility	132.8	131.3	132.0	4.383	127.9	3.808	119.6	0.216	110.3
	Udder	105.2	104.2	106.8	3.658	99.6	2.625	96.5	0.149	90.3

¹54K = ~54,000 markers; 777K = ~777,000 markers; imp = imputed.

ances from the Bayesian mixture model. These variances were estimated based on the DRP derived from the EBV for which a Nordic standardization procedure (<http://www.landbrugsinfo.dk/Kvaeg/Avl/Sider/principles.pdf>) was applied. Therefore, the scales of these variances were different from the original scales of the traits. The additive genetic variances estimated using 54K and 777K marker data were similar in both breeds.

The SNP-effect variances (σ_{v0}^2 and σ_{v1}^2) were dependent on the number of markers (m); the larger the number of markers, the smaller the variance. It was observed that the posterior proportions of SNP in the 2 distributions were similar to the priors. According to the estimated variances in Table 4 and the corresponding prior $\pi = 0.20$, the value of $m[\pi\sigma_{v1}^2 + (1-\pi)\sigma_{v0}^2]$ was similar to the additive genetic variance estimated from the GBLUP model. Among the traits, 89 to 97% of additive genetic variance was accounted for by 20% of the markers in the 54K data or by 2% of the markers in the 777K data.

Genomic Prediction in Nordic Holstein

Reliabilities of genomic predictions for Holstein based on the 54K and HD markers using the 2 alternative models are shown in Table 5. The use of HD mark-

ers led to a small increase in 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 54K markers. We observed that the Bayesian mixture model was superior to the GBLUP model, regardless of which marker data set was used. On average, the increase of reliability using the Bayesian mixture model was 0.5%. On the other hand, imputation of missing genotypes in the 54K data did not yield any improvement of reliability of DGV.

A necessary condition for unbiased genomic prediction is that the regression coefficient of DRP on genomic prediction is 1. As shown in Table 6, using HD markers led to less biased DGV for protein and fertility but not for udder health. Compared with the GBLUP model, the Bayesian model did not reduce bias of DGV. Imputing missing genotypes in the 54K data slightly increased bias compared with the unimputed 54K data.

Genomic Prediction in Nordic RDC

The influences of models and marker data sets on reliability of DGV in RDC (Table 7) were somewhat different from those in Holstein. Imputing missing genotypes in the 54K data improved reliability of DGV for protein, but not for the other 2 traits. The Bayesian mixture model gave very similar reliability as GBLUP,

Table 5. Reliability of direct genomic values using genomic BLUP (GBLUP) and Bayesian mixture based on 54K (~54,000 markers) and 777K (~777,000 markers) data, for Holstein bulls in test data¹

Trait	GBLUP			Bayesian mixture		
	54K	54K _{imp}	777K	54K ($\pi = 0.2$)	54K _{imp} ($\pi = 0.2$)	777K ($\pi = 0.02$)
Protein	0.425	0.426	0.429	0.435	0.434	0.440
Fertility	0.404	0.403	0.413	0.406	0.406	0.416
Udder health	0.370	0.372	0.370	0.375	0.376	0.376
Average	0.400	0.400	0.404	0.405	0.405	0.410

¹Imp = imputed; π = proportion of SNP having large effects.

Table 6. Regression of deregressed proofs on direct genomic values using genomic BLUP (GBLUP) and Bayesian mixture based on 54K (~54,000 markers) and 777K (~777,000 markers) data, for Holstein bulls in test data¹

Trait	GBLUP			Bayesian mixture		
	54K	54K _{imp}	777K	54K ($\pi = 0.2$)	54K _{imp} ($\pi = 0.2$)	777K ($\pi = 0.02$)
Protein	0.853	0.847	0.863	0.855	0.845	0.862
Fertility	0.972	0.963	0.994	0.968	0.958	0.996
Udder health	0.952	0.933	0.946	0.948	0.927	0.946
Average	0.926	0.914	0.934	0.924	0.910	0.935

¹Imp = imputed; π = proportion of SNP having large effects.

based on the 54K 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 unimputed 54K markers, and 0.7% higher than using the imputed 54K markers. When applying the Bayesian mixture model, the increase in reliability using the HD markers was 1.20 and 0.80%, respectively, compared with the unimputed 54K and the imputed 54K markers.

The regression coefficients of DRP on DGV (Table 8) were closer to 1 when DGV were predicted based on the HD markers, indicating a reduction of bias using HD markers. As in Holstein, in RDC the Bayesian mixture model did not reduce bias of DGV, regardless of the marker data set used. In contrast to Holstein, imputing missing genotypes in the 54K data reduced bias of DGV, mainly for protein.

DISCUSSION

This study investigated the advantage of using HD markers for genomic prediction. Based on the present data and models, when going from 54K to HD markers the increase in reliability of DGV was, on average, 0.5% for Holstein and 1.0% for RDC. In addition, genomic predictions were less biased when based on HD markers. The results are consistent with simulation studies assuming a large number of genes affecting the trait.

The study by VanRaden et al. (2011) reported that increasing the number of markers from 54,000 to 500,000 yielded a gain of 1.6% in their simulation study, and the gains were 0.9 and 1.2% using 2 sets of imputed HD marker. Harris and Johnson (2010b) showed very little gain when the number of markers was increased from 20,000 to 1,000,000 in a simulation study.

The Nordic RDC in this study included the Finnish Ayrshire, Swedish Red, and Danish Red populations. The gain in reliability of genomic prediction using the HD markers was larger in RDC than in Holstein. This supports the hypothesis that HD markers give more benefit for genomic prediction across populations than within populations (Toosi et al., 2010). Previous studies on LD and persistence of LD phase (Gautier et al., 2007; de Roos et al., 2008; Villa-Angulo et al., 2009) suggested that genomic selection across populations and breeds requires a higher density of markers than genomic selection within population and breed. With increasing marker density from 54K to 777K, the relative increase of LD (calculated as LD_{777K}/LD_{54K}) was larger for RDC than for Holstein (Table 1). This may explain why RDC obtained a relatively larger gain from HD markers than Holstein.

The number of markers in the HD data set after editing was 11 times the number in the 54K data set. Average pair-wise LD between adjacent markers in HD data set was 3 times as high as in the 54K data set for

Table 7. Reliability of direct genomic values using genomic BLUP (GBLUP) and Bayesian mixture based on 54K (~54,000 markers) and 777K (~777,000 markers) marker data, for Red Dairy Cattle bulls in test data¹

Trait	GBLUP			Bayesian mixture		
	54K	54K _{imp}	777K	54K ($\pi = 0.2$)	54K _{imp} ($\pi = 0.2$)	777K ($\pi = 0.02$)
Protein	0.346	0.358	0.358	0.346	0.357	0.359
Fertility	0.297	0.293	0.304	0.299	0.296	0.307
Udder health	0.244	0.246	0.257	0.243	0.248	0.259
Average	0.296	0.299	0.306	0.296	0.300	0.308

¹Imp = imputed; π = proportion of SNP having large effects.

Table 8. Regression of deregressed proofs on direct genomic values using genomic BLUP (GBLUP) and Bayesian mixture based on 54K (~54,000 markers) and 777K (~777,000 markers) marker data, for Red Dairy Cattle bulls in test data¹

Trait	GBLUP			Bayesian mixture		
	54K	54K _{imp}	777K	54K ($\pi = 0.2$)	54K _{imp} ($\pi = 0.2$)	777K ($\pi = 0.02$)
Protein	0.849	0.875	0.877	0.835	0.864	0.877
Fertility	0.934	0.939	0.980	0.933	0.940	0.980
Udder health	0.851	0.854	0.872	0.839	0.846	0.870
Average	0.878	0.889	0.910	0.869	0.883	0.909

¹Imp = imputed; π = proportion of SNP having large effects.

RDC and 2.7 times for Holstein. Assuming that the same pattern applies to LD between markers and QTL, this suggests much stronger LD between HD markers and genes affecting the trait of interest. Therefore, it was expected that the HD markers would lead to much better genomic predictions. However, the current study shows that the gain from the increased density of the HD markers was small. Several possible reasons exist for this. First, the advantage of increasing LD by HD markers might be counteracted by increasing the number of unknown parameters to be estimated. In the present study, to reduce the number of unknown parameters, the markers in complete LD with the other markers in the data were considered as noninformative markers and thus were deleted. It may be necessary to further reduce the number of markers by deleting the markers that are nearly noninformative. Second, the models used in this study may not be optimal. The results from the current study show that the Bayesian mixture model with 2 normal distributions had a small advantage over the GBLUP model based on the Holstein data. More sophisticated variable selection methods and models would be beneficial for exploiting the potential advantage of HD markers for genomic prediction; for example, mixture models with more than 2 distributions, models using preselected and well-constructed haplotypes or SNP blocks, and models with appropriate weights for different haplotypes or SNP blocks. Third, the HD marker genotypes were, for most of the bulls, not real marker genotypes using HD chips, but imputed ones. Previous studies on imputation from 3,000 to 54,000 marker data have reported that a small imputation allele error rate leads to a substantial loss of prediction reliability, even when only validation animals are imputed and reference animals have real 54K genotypes. Averaged over the results from French, Nordic, and German validations (Chen et al., 2011; Dasonneville et al., 2011), each 1% of imputation allele error rate resulted in a loss of reliability of 1.3 percentage points. It should be also noted

that this study analyzed only 3 traits. The benefits from HD markers may be larger for some traits, such as those traits affected by fewer genes.

Although sizes of reference populations in RDC and Holstein were similar, RDC had lower reliabilities of DGV than Holstein. Average pair-wise LD between adjacent markers was higher in Holstein than in RDC. This indicates that the genetic similarity between individuals in the Holstein population is higher than that in the RDC population, and consequently leads to a higher reliability of genomic predictions in the Holstein population. A previous study (Goddard, 2009) has shown that reliability of genomic prediction depends on the effective population size. Further study is needed on the effective population sizes of current Nordic Holstein and RDC populations.

Several previous studies based on 54K marker data have reported that linear mixed models assuming that effects of all SNP are normally distributed with equal variances perform as well as variable selection models for most traits in dairy cattle (Hayes et al., 2009a; VanRaden et al., 2009). However, for traits having known major genes such as fat percentage, variable selection models are superior over linear mixed models (Cole et al., 2009; Legarra et al., 2011). In the present study, the Bayesian mixture model yielded 0.5% higher reliability than the GBLUP in Holstein, but the advantage of the mixture model was not observed in RDC, regardless of the marker data used. This contradicts the expectation that a variable selection model would have a greater advantage over a GBLUP model when using HD marker data than when using 54K marker data. At least 3 possible reasons could explain this. First, the mixture model with 2 distributions may not be an optimal model to describe actual distribution of SNP effects. Second, the mixture model may be more sensitive to imputation errors than the GBLUP model. Third, the data information may not be sufficient to efficiently distinguish the SNP with large effects from those with small effects.

Using the GBLUP model, the number of the mixed models equations is not determined by the number of markers, but by the number of individuals. Therefore, the computational demand is almost the same when using the 54K or HD data. Using the Bayesian mixture model, the number of equations is determined by the number of markers. Consequently, the computing time increases with increasing the number of markers. For the analysis of Holstein data in our computing system (Intel Xeon 2.93 GHz processor), given the inverted $\mathbf{G}$ matrix, the GBLUP model took less than 10 min per trait. It took about 6 min to build the $\mathbf{G}$ matrix and calculate the inverted $\mathbf{G}$ matrix based on the 54K marker data, and about 50 min based on the HD data. The Bayesian mixture model with Gibbs sampling approach (total 50,000 samples) took about 10 h when using the 54K data, and about 120 h when using the HD data.

Imputation of missing genotypes in 54K marker data is expected to improve genomic predictions. However, the imputation procedure used in this study to infer missing genotypes in the 54K data did not improve genomic predictions, except for protein in RDC. In the analysis based on the 54K data with missing genotypes, the missing individual genotypes were replaced with population expectations. Thus, individuals with missing genotypes of a particular marker did not contribute to the estimated effect of this marker, and the DGV of the individual did not include the effect of this marker. Replacing missing genotypes with population expectations was a simple imputation. In the current data, there were only about 4% missing genotypes in the 54K marker data. With the small proportion of missing genotypes, superiority of a good imputation procedure over a simple imputation procedure could be less important. This might partly explain why inferring missing individual marker genotypes in the 54K data using a sophisticated imputation procedure did not lead to a clear improvement of genomic prediction, compared with replacing missing genotypes with population expectations.

In conclusion, HD marker data have the potential to increase reliability of genomic predictions. However, the gain of genomic predictions using HD markers is small, based on current data and models. Further studies are needed to exploit the potential advantage of HD markers in genomic predictions.

ACKNOWLEDGMENTS

We thank the Danish Cattle Federation (Aarhus, Denmark), Faba Co-op (Helsinki, Finland), Swedish Dairy Association (Stockholm, Sweden), and Nordic Cattle Genetic Evaluation (Aarhus, Denmark) for pro-

viding data. This work was performed in the project “Genomic Selection—From function to efficient utilization in cattle breeding (grant no. 3405-10-0137),” funded under Green Development and Demonstration Programme by the Danish Directorate for Food, Fisheries and Agri Business (Copenhagen, Denmark), the Milk Levy Fund (Aarhus, Denmark), VikingGenetics (Randers, Denmark), Nordic Cattle Genetic Evaluation (Aarhus, Denmark), and Aarhus University (Aarhus, Denmark).

REFERENCES

- Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93:743–752.
- Browning, B. L., and S. R. Browning. 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.
- Chen, J., Z. Liu, F. Reinhardt, and R. Reents. 2011. Reliability of genomic prediction using imputed genotypes for German Holsteins: Illumina 3K to 54K bovine chip. The 2011 Interbull Open Meeting, Stavanger, Norway. Interbull, Uppsala, Sweden.
- Cole, J. B., P. M. VanRaden, J. R. O’Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggins. 2009. Distribution and location of genetic effects for dairy traits. *J. Dairy Sci.* 92:2931–2946.
- Dassonneville, R., R. F. Brondum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M. S. Lund, V. Ducrocq, and G. Su. 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. *J. Dairy Sci.* 94:3679–3686.
- de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. *Genetics* 183:1545–1553.
- de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 179:1503–1512.
- Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* 43:1.
- Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, C. Grohs, A. Boland, J. G. Garnier, D. Boichard, G. M. Lathrop, I. G. Gut, and A. Eggen. 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics* 177:1059–1070.
- Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: Application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal* 1:21–28.
- Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* 136:245–257.
- Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. *Genetics* 182:343–353.
- Harris, B. L., and D. L. Johnson. 2010a. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J. Dairy Sci.* 93:1243–1252.
- Harris, B. L., and D. L. Johnson. 2010b. The impact of high density SNP chips on genomic evaluation in dairy cattle. *Interbull Bull.* 42:40–43.
- Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92:433–443.
- Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. (Camb.)* 91:47–60.

- Legarra, A., C. Robert-Granie, P. Croiseau, F. Guillaume, and S. Fritz. 2011. Improved Lasso for genomic selection. *Genet. Res. (Camb.)* 93:77–87.
- Liu, Z. T., F. R. Seefried, F. Reinhardt, S. Rensing, G. Thaller, and R. Reents. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genet. Sel. Evol.* 43:19.
- Lund, M. S., S. P. de Ross, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genet. Sel. Evol.* 43:43.
- Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* 4:e5350.
- Matukumalli, L. K., S. Schroeder, S. K. DeNise, T. Sonstegard, C. T. Lawley, N. Georges, W. Coppieters, K. Gietzen, J. F. Medrano, G. Rincon, D. Lince, A. Eggen, L. Glaser, G. Cam, and C. Van Tassell. 2011. Analyzing LD blocks and CNV segments in cattle: Novel genomic features identified using the BovineHD BeadChip. *Pub. No. 370-2011-002*. Illumina Inc., San Diego, CA.
- Meuwissen, T., and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185:623–631.
- Meuwissen, T. H. E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* 41:35.
- Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2008. Genomic selection using different marker types and densities. *J. Anim. Sci.* 86:2447–2454.
- Su, G., B. Guldbrandtsen, V. R. Gregersen, and M. S. Lund. 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. *J. Dairy Sci.* 93:1175–1183.
- Su, G., P. Madsen, U. S. Nielsen, E. A. Mäntysaari, G. P. Aamand, O. F. Christensen, and M. S. Lund. 2012. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. *J. Dairy Sci.* 95:909–917.
- Toosi, A., R. L. Fernando, and J. C. M. Dekkers. 2010. Genomic selection in admixed and crossbred populations. *J. Anim. Sci.* 88:32–46.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423.
- VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* 43:10.
- VanRaden, P. M., and P. G. Sullivan. 2010. International genomic evaluation methods for dairy cattle. *Genet. Sel. Evol.* 42:7.
- VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92:16–24.
- Villa-Angulo, R., L. K. Matukumalli, C. A. Gill, J. Choi, C. P. Van Tassell, and J. J. Grefenstette. 2009. High-resolution haplotype block structure in the cattle genome. *BMC Genet.* 10:19.
- Weigel, K. A., G. de los Campos, O. Gonzalez-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J. Dairy Sci.* 92:5248–5257.