Improving genomic prediction for Danish Jersey using a joint Danish-US reference population

G. Su[1], U. S. Nielsen[2], G. Wiggans[4], G. P. Aamand[3], B. Guldbrandtsen[1], M. S. Lund[1]
[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark, [2]Knowledge Center for Agriculture, Aarhus, Denmark, [3]Nordic Cattle Genetic Evaluation, Aarhus, Denmark, [4]Agricultural Research Service, USDA, Beltsville, MD

ABSTRACT: Accuracy of genomic prediction depends on the information in the reference population. Achieving an adequate sized reference population is a challenge for genomic prediction in small cattle populations. One way to increase the size of reference population is to combine reference data from different populations. The objective of this study was to assess the gain of genomic prediction accuracy when including US Jersey bulls in the Danish Jersey reference population. The data included 1,262 Danish progeny-tested bulls and 1,157 US progeny-tested bulls. Genomic breeding values (GEBV) were predicted using a GBLUP model from the Danish reference population and the joint Danish-US reference population. The traits in the analysis were milk yield, fat yield, protein yield, fertility, mastitis, longevity, body conformation, feet & legs, and longevity. Eight of the nine traits benefitted from the inclusion of US Jersey bulls in the reference population. The gains ranged from 1.6% points for fertility to 12.5% points for udder conformation. The exception was longevity for which the joint reference population resulted in a loss of 5.5% points in reliability of GEBV. Averaged over all nine traits, reliability of GEBV using the joint reference population was 4.0% points higher than the reliability of GEBV using the Danish reference population alone. The results confirm that exchanging reference data to increase the size of reference population is an efficient approach to increase the accuracy of genomic prediction, especially for the populations with small number of reference bulls.

Keywords:
genomic selection
reliability
reference population

Introduction

One of the most important factors affecting accuracy of genomic prediction is the size of the reference population (Daetwyler et al., 2008; Goddard, 2009). In dairy cattle, reference populations are usually composed of progeny-tested bulls, since they have reliable phenotypic information from a large group of daughters. However, for a single country, the number of progeny-tested bulls is limited, especially for small dairy cattle populations. Therefore, achieving an adequate sized reference population is a challenge for genomic selection in small populations. One approach to increase the size of reference population is to combine reference populations from various countries. This approach has been shown to be very efficient (Schenkel et al., 2009; Jorjani et al., 2011; Lund et al., 2011; VanRaden et al., 2012; Zhou et al., 2013).

Danish Jersey is a small dairy cattle population. Currently, only about 1,200 – 1,400 progeny-tested bulls (depending on trait) are available as reference bulls. Due to the small reference population, accuracy of genomic prediction in the Danish Jersey is much lower than those in the Danish Holstein and Red Cattle populations. In 2013, marker data for Danish and US Jersey bulls were exchanged, creating a joint reference population for genomic prediction of Danish Jersey. The objective of this study was to assess the gain of genomic prediction accuracy for Danish Jerseys using the joint Danish-US Jersey reference population.

Materials and Methods

Data. The data in the analysis included 1,262 Danish Jersey bulls and 1,157 US Jersey bulls. Danish Jersey bulls were genotyped with Illumina Bovine SNP50 chip. US Jersey bulls were genotyped either with the standard Illumina Bovine SNP50 chip or the GeneSeek Genomic Profiler chip with 76,999 SNP). The markers which are not in the Bovine SNP50 chip were excluded in the Danish prediction. Sporadic missing genotypes were imputed using Beagle (Browning and Browning, 2009). After removing markers with allele frequency less than 1%, 41,723 autosomal markers were used to predict genomic breeding values.

The Danish data were divided into reference data and validation data by birth date. The bulls born before 2005 were used as reference bulls. This led to about 20% Danish bulls being validation bulls. All the US bulls were included in a joint Danish-US reference population. Deregressed proofs (DRP) were derived from the published EBV in the Interbull 2014-01 evaluation and used as response variable in the analysis. The published EBV were standardized to a mean of 100 and a standard deviation of 10 in a particular year. In the validation data, bulls were required to have DRP with a minimum reliability of 0.20. In total, nine combined traits were analyzed. The number of bulls in the Danish reference data, the joint reference data and the validation data are shown in Table 1.

Statistical model. Breeding values were predicted using conventional BLUP model based on the Danish reference data, GBLUP model based on the Danish reference data and the joint reference data. The GBLUP model is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}_g\mathbf{g} + \mathbf{Z}_u\mathbf{u} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of DRP, μ is the overall mean, $\mathbf{1}$ is a vector of ones, $\mathbf{g}$ is the vector of additive genomic effects, $\mathbf{u}$ is the vector of residual polygenic effects, $\mathbf{Z_g}$ and $\mathbf{Z_u}$ are incidence matrices linking $\mathbf{g}$ and $\mathbf{u}$ to $\mathbf{y}$, and $\mathbf{e}$ is the vector of residuals.

It is assumed that $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma^2_g)$, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma^2_u)$, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D}\sigma^2_e)$, where **G** is a genomic relationship matrix (VanRaden, 2008; Hayes et al., 2009), σ^2_g is the genomic additive genetic variance, **A** is a pedigree-based relationship matrix, σ^2_u is the residual additive genetic variance, **D** is a diagonal matrix and σ^2_e is the residual variance. **D** has diagonal elements $d_{ii} = (1-r^2_{DRP})/r^2_{DRP}$, and is used to account for heterogeneous residual variances due to different reliabilities of DRP (r^2_{DRP}). Additive genetic and residual variances used for predictions were those used in Nordic routine genetic evaluation. In this analysis, 80% of total additive genetic effect was assigned to σ^2_g, and the remaining 20% to σ^2_e. Genomic breeding value (GEBV) was calculated as GEBV= $\hat{g} + \hat{u}$.

Validation of genomic predictions. Reliability of predicted breeding values was calculated as the squared correlation between predicted breeding values and DRP and then divided by reliability of DRP for the validation animals. To account for possible underestimation of the validation reliability due to the validation bulls comprising selected bulls, an adjusted reliability was calculated as validated reliability of genomic prediction plus the difference between expected (according to prediction error variance obtained from mixed model equation) and validated reliabilities of pedigree index of the validation bulls. Bias of genomic predictions was measured using regression of DRP on GEBV.

Results and Discussion

Genomic prediction using Danish reference population. As shown in Table 2, validation reliability (based on correlation between prediction and DRP) of the pedigree index ranged from 0.034 to 0.362 with an average of 0.145, while expected reliability (based on inversed coefficient matrix of mixed model equation) was between 0.103 and 0.303 with an average of 0.254. The lower validation reliabilities could be because the validation animals were progeny-tested bulls which were preselected according to parent average breeding value. This reduces the correlation between predictions and DRP, and consequently the validation reliability.

Despite a small reference population of Danish Jersey, reliability of genomic prediction was much higher than conventional pedigree index (based on sire and maternal grandsire information in this study). Averaged over nine traits, the reliability of GEBV was 11.6% points higher than the pedigree index. It was observed that GEBV was inflated, because the regression coefficient was less than one for all the traits (Table 3), most seriously for feet & legs.

Gain in prediction reliability with a joint reference population. Eight of the nine traits benefitted from the joint reference population. The gains by including US Jersey bulls in the reference population ranged from 1.6% points for fertility to 12.5% points for udder conformation (Table 2). For longevity there was a loss of 5.6% points in reliability of GEBV. Averaged over all nine traits, reliability of GEBV using the joint reference population was 4.0% points higher than the reliability of GEBV using the Danish reference population alone. However, the joint reference population did not reduce the inflation of GEBV (Table 3).

Inclusion of 1,157 US Jersey bulls resulted in a large increase in accuracy of genomic prediction. At least two effects explain this large gain. First, the Danish reference population was small; the inclusion of 1,157 US Jersey bulls doubled the size of reference population. The large benefit from including reference animals of another population to a small national reference population has been reported by Zhou et al. (2013), who added 4,400 Danish progeny-tested Holstein bulls to the Chinese reference population which comprised 1,500 Chinese Holstein cows. The gain from the inclusion of Danish bulls was 29% points for Chinese Holstein bulls and 7% points for Chinese Holstein cows.

Secondly, there are strong genetic links between Danish Jersey and US Jersey. Semen of US Jerseys has been used in the Danish Jersey population for a long time, especially during the period from 1985 to 1995. Today the US Jersey breed proportion is about 38% in the Danish Jersey population (http://www.vikinggenetics.com.au /breeds/viking-jersey/about-viking-jersey). The importance of relationship between populations that are combined for genomic prediction across populations has been reported in many previous studies (Brøndum et al., 2011; Zhou et al., 2013; Zhou et al., 2014). .

The size of reference population in US Jersey was moderate large, the inclusion of Danish Jersey bulls gave relatively small increase in size of reference population. The gain in accuracy of genomic prediction for US Jersey by including Danish Jersey in US reference population was 1.8% points, averaged over the traits under analyses (https://www.cdcb.us/reference/changes/eval1401.htm).

Conclusion

The inclusion of 1,157 US Jersey bulls doubled the size of Danish Jersey reference population, and led to a large improvement in accuracy of genomic prediction for Danish Jerseys. The results indicate that exchanging reference data to increase size of reference population is an efficient approach to increase accuracy of genomic selection, especially for populations with a small number of reference bulls.

Literature Cited

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

Browning, B. L. and Browning, S. R. (2009). Am J Hum Genet 84:210-223.

Daetwyler, H. D., Villanueva, B. and Woolliams, J. A. (2008). PLoS ONE 3:e3395.

Goddard, M. E. (2009). Genetica 136:245-257.

Hayes, B. J., Visscher, P. M. and Goddard, M. E. (2009). Genetics Research 91:47-60.

Jorjani, H., Jakobsen, J., Nilforooshan, M. A. et al (2011). Interbull Bulletin 43:5-8.

Lund, M. S., de Ross, S. P., de Vries, A. G. et al. (2011). Genet Sel Evol 43:43.

Schenkel, F. S., Sargolzaei, M., Kistemaker, G. et al. (2009). Interbull Bulletin 39:51-57.

VanRaden, P. M. 2008. J Dairy Sci 91:4414-4423.

VanRaden, P. M., Olson, K. M. , Null, D. J. et al. (2012) Interbull Bulletin 46:75-79.

Zhou, L., Ding, X., Zhang, Q. et al. (2013). Genet Sel Evol 45:7.

Zhou, L., Heringstad, B., G. et al. (2014) J Dairy Sci In press.

Table 1. Number of bulls in the reference and validation data

Trait	Validation	Ref_DK	REF_ DK+US	Trait	Validation	Ref_DK	REF- DK+US
Milk	233	1027	2184	Longevity	150	1006	2073
Fat	233	1027	2184	Body conf.	213	1018	2002
Protein	233	1027	2184	Feet & legs	225	1019	1999
Fertility	221	1024	2145	Udder conf.	225	1019	2001
Mastitis	234	1027	2143				

Table 2. Reliability of pedigree index (PI) and genomic prediction using Danish reference population (DK) or joint reference population (DK+US)

Trait	R^2_{exp} (1)	R^2_{valid} (2)			R^2_{adjust} (3)		Dif
	PI*	PI	GEBV DK	GEBV DK+US	GEBV DK	GEBV DK+US	R^2_{DK+US} - R^2_{DK}
Milk	0.297	0.156	0.360	0.415	0.501	0.555	0.054
Fat	0.297	0.034	0.198	0.234	0.461	0.497	0.036
Protein	0.303	0.098	0.272	0.311	0.477	0.516	0.039
Fertility	0.103	0.071	0.196	0.212	0.227	0.243	0.016
Mastitis	0.292	0.362	0.415	0.441	0.344	0.371	0.027
Longevity	0.280	0.112	0.216	0.160	0.384	0.328	-0.056
Body conf.	0.224	0.251	0.369	0.404	0.342	0.376	0.034
Feet & legs	0.238	0.078	0.103	0.183	0.263	0.343	0.080
Udder conf.	0.251	0.142	0.216	0.341	0.325	0.450	0.125
Mean	**0.254**	**0.145**	**0.261**	**0.300**	**0.369**	**0.409**	**0.040**

*Pedigree index for validation bulls, calculated from the Danish reference data. This corresponded to the pedigree index calculated from sire and maternal grandsire.

(1): Reliability of PI estimated from inversed coefficient matrix of mixed model equation.

(2): Reliability calculated as squared correlation between prediction and deregressed proof (DRP) and then divided by reliability of DRP.

(3): Reliability of GEBV calculated as $R^2_{adjust} = R^2_{valid} + (R^2_{exp(PI)} - R^2_{valid(PI)})$

Table 3. Regression coefficient of deregressed proof on genomic prediction using the Danish (DK) or joint reference population (DK+US).

Trait	DK	DK+US	Trait	DK	DK+US
Milk	0.891	0.830	Longevity	0.902	0.720
Fat	0.701	0.707	Body conf.	0.898	0.871
Protein	0.750	0.713	Feet & legs	0.592	0.701
Fertility	0.949	0.918	Udder conf.	0.762	0.928
Mastitis	0.898	0.916	**Mean**	**0.816**	**0.812**