Dairy Cattle Genetic Evaluation Using Genomic Information

Z. Liu, F. Seefried, F. Reinhardt and R. Reents
vit *w.V., Heideweg 1, 27283 Verden, Germany*

Abstract

Deterministic genomic models, including a high number of random marker effects and a polygenic effect, were presented for genomic evaluation of dairy cattle. The genomic models covered a linear BLUP model assuming equal variance for all markers and five non-linear genomic models allowing variable marker variances. Marker and phenotypic data were obtained from a workshop for investigating the alternative genomic models. A polygenic model was fitted to the simulated data additionally, in order to compare conventional and genomic evaluations. Although the genomic models differ in estimates of marker effects, all of them resulted in relatively high correlation between true and estimated genomic breeding values for the training data set, ranging from 0.85 to 0.90. For the validation data set, difference in the correlation increased among the genomic models. The polygenic model was clearly the worst model, with lowest correlation between true and estimated genomic breeding values, particularly for the validation set. The ranking of the genomic models may change for real data. A brief status report was given on the German genome project GenoTrack.

1. Introduction

Genomic selection (Meuwissen *et al.,* 2001) based on massive marker information, e.g. single nucleotide polymorphism (SNP), can increase accuracy of pre-selection of breeding animals significantly. Worldwide application of genomic selection has started recently (Daetwyler *et al.,* 2007; De Roos *et al.,* 2007; Ducrocq, 2008; Harris *et al.,* 2008; VanRaden, 2008). The objectives of this study were to develop a genomic enhanced genetic evaluation system for dairy cattle, and to compare alternative genomic evaluation models using a simulated data set.

2. Materials and Methods

2.1. Data materials from a simulation study

Marker and phenotypic data were obtained from a QTL-MAS Workshop in 2008 (Lund *et al.,* 2008). A total of 4665 animals were genotyped for 6000 bi-allelic markers that were evenly located on six chromosomes. The genotyped animals from the training data set came from four discrete generations with 165 animals in generation 0 and 1500 animals in generations 1 to 3 each. Model for generating phenotypic records included, besides a random error effect, 48 QTLs; and none of them were located on chromosome 6. 16 biggest QTLs explained 96.3% total genetic variance. A validation data set contained three generations, generations 4 to 6, of genotyped animals without phenotypic records, 400 animals per generation. A swine-type pedigree structure was chosen in the simulation, 45 sires were randomly mated to 10 dams to generate 10 progeny each, resulting 45 sires (450 dams) with 100 (10) genotyped progeny each.

2.2. Genomic evaluation models

A statistical model was applied to both genotypic and phenotypic data of genotyped animals:

$$q_i = \mu + a_i + \sum_{j=1}^{p} z_{ij} u_j + e_i$$

where q_i is a deregressed proof (DPRF) or daughter yield deviation (DYD) of a bull *i* or a yield deviation (YD) of a cow *i*, for the simulated data set q_i is the trait value, μ is a general mean, a_i is polygenic effect of animal *i*, p is number of fitted bi-allelic markers ($j = 1, \cdots, p$), z_{ij} is genotype value (-1 and 1 for two homozygotes and 0 for heterozygote) of marker j of animal *i*, u_j is random

regression coefficient for marker *j*, and e_i is residual effect for the record of animal *i*. A small fraction of genetic variance, usually not more than 1%, was assumed for modelling the polygenic effect. Fitting a polygenic effect can avoid the problem that the markers captured the relationship among animals if the genomic model did not include the polygenic effect (Habier *et al.*, 2007), and thus it can make the estimates of marker effects more persistent over generations (Solberg *et al.*, 2008). Because DYDs resulting from a multi-trait model, e.g. random regression test-day model (Liu *et al.*, 2004), cannot be optimally analysed with the single trait genomic model, a single trait deregression procedure using full animal pedigree was applied to derive DPRFs for genotyped bulls. Effective daughter contribution (EDC) was used as weighting factor for DPRF in the genomic evaluation. For the simulated data, no deregression or calculation of DYD was required, because there were no other effects generated in the simulation.

All markers from the simulated data set were considered simultaneously in the genomic evaluation. Alternative genomic models differed in prior variance functions of the fitted markers. When all markers were assumed to have equal variance, denoted as linear model EQ, a linear regression of marker phenotypic deviation on genetic effect was expected (VanRaden, 2008). In contrast to the linear BLUP model EQ, variances of markers were assumed to be a function of marker effect estimates. Under such a genomic model, the regression of marker phenotype on genetic effect was non-linear (VanRaden, 2008). In general, the non-linear models regressed small marker effects more strongly towards zero than the linear BLUP genomic model EQ, and at the same time the non-linear models allowed bigger markers having bigger variances. In statistics, such models with weights or variances depending on effect estimates themselves were termed as robust regression model (Draper and Smith, 1998). Table 1 describes the genomic models tested using the simulated data set from the 2008 QTL-MAS workshop. The two submodels with an exponential function, E1 and E2, resembled the US genomic evaluation model (VanRaden, 2008).

The fitted polygenic effect of the genomic models was analysed in the same way as in conventional genetic evaluation, i.e. using full pedigree and identical grouping of phantom parent groups. The identical modelling of the polygenic effect ensured that the polygenic estimated breeding values (EBV) were as close as possible to those from conventional evaluations.

2.3. Estimating marker allele effects

Mixed model equations of the genomic and polygenic models were solved with a special Gauss-Seidel algorithm (Lagarra and Misztal, 2008), which was developed for statistical models containing a high number of effects. Application of the computing algorithm required that residuals must be updated periodically for avoiding accumulation of rounding errors. Prior marker effect estimates can be used for reducing computing time; for the non-linear models prior marker variances were also needed in addition to the prior marker effect estimates. Marker allele frequencies were estimated using the gene content method (Gengler *et al.*, 2007) for the base population. A single phantom parent group was currently formed to estimate the base population allele frequency, though more groups could be considered. To obtain reliability of direct genomic breeding values (DGV), realised genomic relationship matrix was set up using the estimated base population allele frequencies and marker genotypes (VanRaden, 2008). Estimated DGV were then combined with conventional EBV for genotyped animals in routine genomic evaluation, weighted by their respective reliabilities, to obtain final genomic EBV (GEBV) and associated reliability with the standard selection index approach. For the simulated data, final GEBV was the sum of estimated DGV and polygenic EBV.

Table 1. Alternative genomic models with prior marker variance functions (σ_j^2 is variance of marker j, $\bar{\sigma}^2$ is average marker variance, and $\hat{s}_j$ is standardised estimate of marker j.)

Variants of genomic model	Marker variance function
Linear BLUP model with equal variances (EQ)	$\sigma_j^2 = \bar{\sigma}^2 \times 1$
Non-linear models with variance functions:	
Exponential function (E1)	$\sigma_j^2 = \bar{\sigma}^2 \times 1.12^{\lvert\hat{s}_j\rvert}$
Exponential function (E2)	$\sigma_j^2 = \bar{\sigma}^2 \times 1.25^{\lvert\hat{s}_j\rvert}$
Linear function (LW)	$\sigma_j^2 = \bar{\sigma}^2 \times \lvert\hat{s}_j\rvert$
Quadratic function (Q1)	$\sigma_j^2 = \bar{\sigma}^2 \times \hat{s}_j^2$
Quadratic function with effect limits (Q2)	$\sigma_j^2 = \bar{\sigma}^2 \times \hat{s}_j^2$ with lower/upper limits on $\hat{s}_j$
Polygenic model without markers (PG)	$\sigma_j^2 \approx \bar{\sigma}^2 \times 0$

3. Results and Discussion

All computation was conducted on a Linux server using own Fortran 95 programs. Convergence criterion was defined as logarithm of sum of squared difference in solutions between last and current rounds divided by sum of squared solutions from current round. Iteration process was considered to be converged, when the convergence criterion was less than –10. Usage of RAM was limited for the simulated data set containing 4665 genotyped and phenotyped animals. For the non-linear genomic models, 100 rounds of 'burn-in' were executed using equal marker variance, followed by using variable marker variances that depended on marker effect estimates from previous round. Total clock time was between 25 to 120 minutes, depending on the models. The simulated total heritability values was 0.3 and our REML estimate with a polygenic model was 0.304. For all the investigated genomic models, a very low polygenic heritability of 0.001 (and a total marker heritability of 0.303) was used; whereas the polygenic model had a polygenic heritability of 0.303 (a total marker heritability of 0.001). Weighting factor for each phenotypic record was set to 1 in all the analyses.

3.1. Marker effect estimates

Figure 1 shows absolute marker effect estimates, expressed in standard deviation of average marker under the linear BLUP model EQ (left) and the non-linear genomic model LW (right). It can be seen that many small markers had non-zero estimates under the linear genomic model EQ. In contrast, the small marker estimates were practically 0 under the non-linear genomic model LW with the linear marker variance function. Usually, a QTL was signalled by several closely linked markers with larger effects. All simulated QTLs, except those explaining less than 0.5% genetic variance, were identified by all the genomic models, including the linear genomic model EQ. However, the signals for QTLs were stronger for the non-linear genomic models than for the linear one. Although no QTLs were simulated on chromosome 6, non-zero marker estimates were obtained, notably from the linear genomic model EQ. The two non-linear models Q1 and Q2, which assumed the strongest prior marker variances, identified fewest markers with non-zero effects and no false positives on chromosome 6; whereas the remaining genomic models found more small markers and also some markers with non-zero effect estimates on chromosome 6.

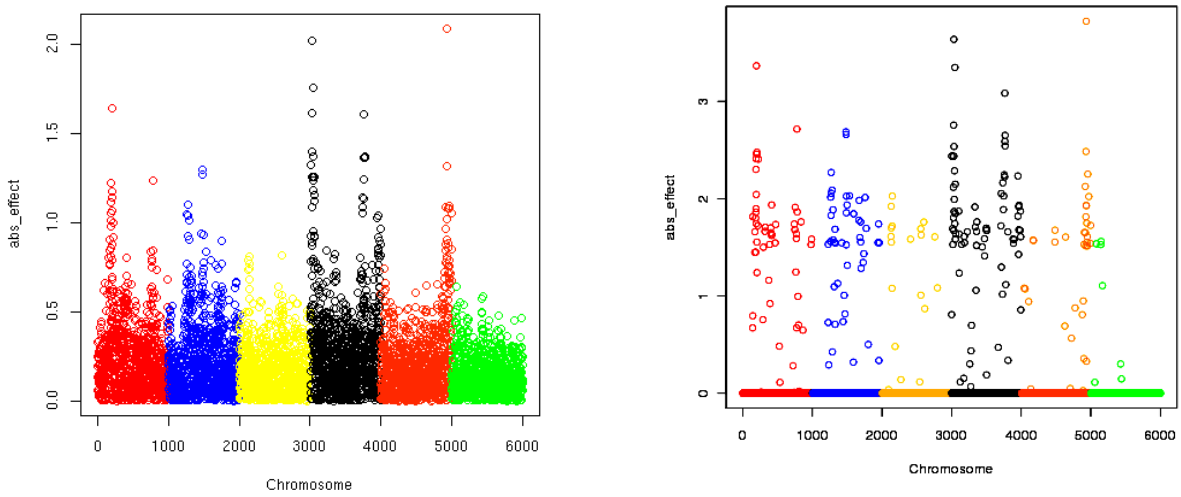

Figure 1. Marker effect estimates of the linear genomic model EQ with equal marker variances (left) and the non-linear genomic model LW with linear variance function (right).

3.2. Estimated genomic breeding values

Because a very low heritability was assumed for all the genomic models for the simulated data, polygenic EBVs from the genomic models were negligible, thus only estimated DGV were analysed further on. No biases were observed in the training data set for all the models. Table 2 gives correlation of estimated DGVs with simulated true genomic breeding values for all the genomic models and the polygenic model. Overall, the correlation was between 0.86 and 0.90 for the training data set for all the genomic models, whereas lower correlation was found for the validation set, as expected. In comparison, the polygenic model resulted in lower correlation with the true genomic breeding values for the training data set, much lower for the validation set due to the fact that the polygenic model did not use genotypic information for predicting breeding values. In contrast to the difference in marker effect estimates, the six genomic models differed less in estimated DGVs.

Table 2. Correlation of estimated genomic breeding values with true breeding values for the genomic and polygenic models.

		No. animals	Genomic models						Polygenic model PG
			EQ	E1	E2	LW	Q1	Q2	
Generation number of training data set	0	165	.88	.88	.87	.85	.86	.82	.76
	1	1500	.87	.87	.88	.87	.87	.83	.67
	2	1500	.90	.90	.92	.90	.91	.88	.75
	3	1500	.87	.88	.90	.89	.90	.87	.69
	All	**4665**	**.88**	**.88**	**.90**	**.88**	**.89**	**.86**	**.70**
Generation number of validation data set	4	400	.76	.76	.79	.80	.85	.79	.25
	5	400	.78	.79	.84	.86	.84	.83	.09
	6	400	.74	.74	.78	.83	.83	.81	.03
	All	**1200**	**.76**	**.77**	**.81**	**.83**	**.84**	**.80**	**.15**

The polygenic model had a correlation of 0.71 between its EBVs and estimated DGVs of the genomic model Q2 for the training set, and the correlation droped to 0.34 for the validation set. These correlations were slightly higher than the correlations between true and estimated breeding values of the polygenic model.

As predictive ability of the models concerned, it seems that all the genomic models gave much higher correlation between true and estimated DGVs than the polygenic model; the genomic models with stronger prior variances, e.g. Q1, Q2 and LW, tended to have higher correlated estimated DGVs with true values than the genomic models with weaker variances, e.g. EQ, E1 and E2. The ranking of the genomic models may be data dependent.

3.3. Estimated genomic breeding values of individual animals

In the training set there were 45 sires with 100 and 450 dams with 10 genotyped progeny each. Table 3 shows correlations between estimated DGVs and true breeding values or EBVs of the polygenic model PG for the sires and dams. All the correlations were quite high for the sires with 100 progeny and also for the dams with 10 progeny. EBVs of the polygenic model had the lowest correlation with true values than estimated DGVs from all the genomic models. The high correlation of the estimated DGVs with true breeding values for the sires or dams can be attributed to high frequencies of their marker haplotypes in the training set. Marker haplotypes of the sires or dams were, therefore, estimated reasonably accurately and so were their genomic breeding values as sum of the haplotypes. In contrast to the variable reliabilities or accuracy of estimated DGV between animals, estimates of all the markers should have equal reliability. But covariances or correlations between estimates of any pair of markers were dependent on the frequencies of the marker haplotypes in the training set.

Table 3. Correlation of estimated genomic breeding values of sires with 100 and dams with 10 genotyped progeny each in training set with true breeding values and EBVs of the polygenic model.

		Genomic models						EBV of
		EQ	E1	E2	LW	Q1	Q2	model PG
45 sires	True BV	.95	.95	.96	.94	.96	.93	.93
	EBV of model PG	.96	.96	.96	.95	.94	.93	
450 dams	True BV	.90	.90	.91	.88	.88	.85	.81
	EBV of model PG	.85	.85	.85	.80	.77	.76	

4. Application to real genotypic and phenotypic data of German Holsteins

For the German national genome project, GenoTrack, (Thaller 2008), a total of 2830 Holstein bulls, born in 1998 throughout 2002, were genotyped using Illumina chip Bovine SNP50 BeadChip. Additionally, about 600 older bulls have being genotyped. By setting a minimum minor allele frequency of 0.01, 45,488 SNP markers remained. Male animals were not expected to be heterozygous for markers located only on sex chromosome X.

As phenotypic record, DPRF of bulls or YD of cows were chosen for the genomic evaluation based on a single trait genomic model, although DYD derived from multi-trait models (Liu *et al.*, 2004) can describe phenotypic information from the conventional multi-trait models more accurately. DPRF were obtained with full animal pedigree using a software from Interbull Centre. A total of 44 traits from seven trait groups were considered in genomic evaluation: milk production (3 traits), somatic cell scores (1 trait), function longevity (1 trait), calving (4 traits), female fertility (6 traits), workability (4 traits) and conformation (25 traits). All bulls with at least 10 EDC were included in the deregression procedure. As cows are going to be genotyped in near future, YD of the cows may be preferred to DPRF for cows, because of possibly extreme DPRF values caused by cows' low reliability values. Estimated marker allele frequencies using gene content method (Gengler *et al.*, 2007) were used to approximate genomic relationship matrix, which were then used to derive reliabilities of DGV estimates. Conventional EBVs from official evaluation were combined with the DGV estimates using their reliabilities as weights to calculate final GEBV. German total merit index RZG and other related indices

were derived on the basis of the combined GEBV. Waiting bulls of German Holstein breed, born from 2003 onwards, are going to be genotyped. A validation of the genomic evaluation system will be conducted using data of these younger bulls.

5. Further developments

Currently, genomic evaluations are based on a single trait model, which has no longer been the standard for conventional genetic evaluations. In order to make optimal use of phenotypic information from conventional, multi-trait model evaluations, an extension of the genomic models to multi-trait evaluations is needed. With GEBV of young animals available, breeders will conduct pre-selection using the GEBV information, which may cause problem for conventional evaluations unless all genotyping information is considered for genetic evaluations. Reliability approximation for estimated direct genomic breeding values needs to be improved to consider ever more genotyped animals and possible overestimation of the reliabilities. Until now, conventional and direct genomic proofs are estimated separately and then combined using selection index theory, which may double count phenotypic information. A joint analysis of polygenic and direct genomic breeding values is a more accurate way to obtain combined genomic breeding values.

6. References

Daetwyler, H.D., Schenkel, F.S., Sargolzaei, M. & Robinson, J.A.B. 2008. *J. Dairy Sci. 91,* 3225.

De Roos, A.P.W., Schrooten, C., Mullaart, E., Calus, M.P.L. & Veerkamp, R.F. 2007. *J. Dairy Sci. 90,* 4821-4829.

Ducrocq, V. 2008. Genomic evaluation and selection in France.

Draper N.R. & Smith, H. 1998. *Applied regression analysis.* Third Edition. John Wiley & Sons, Inc.

Gengler, N., Mayeres, P. & Szydalowski, M. 2007. *Animal 1,* 21-28.

Habier, D., Fernando, R.L. & Dekkers, J.C.M. 2007. *Genetics 177,* 2389.

Harris, B.L., Johnson, D.L. & Spelman, R.J. 2008. *ICAR meeting,* Niagara Falls, USA.

Legarra, A. & Misztal, I. 2008. *J. Dairy Sci. 91,* 260-266.

Liu, Z., Reinhardt, F., Bünger, A. & Reents, R. 2004. *J. Dairy Sci. 87,* 1896-1907.

Lund, M. *et al.* 2008. *QTL-MAS Workshop data simulation.* Uppsala, Sweden, March 2008.

Meuwissen, T.H., Hayes, B.J. & Goddard, M.E. 2001. *Genetics 157,* 1819-1829.

Solberg, T.R. *et al. EAAP 2008,* Vilnius, September 2008.

Thaller *et al.* 2008. German genome research project GenoTrack.

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

7. Acknowledgements

Paul VanRaden, George Wiggans, Sander de Roos, and Nicolas Gengler were thanked for sharing experiences and helpful discussions. German research, breeding and governmental organisations were acknowledged for co-operation and funding the project.