Animal (2014), **8:2**, pp 208–216 © The Animal Consortium 2014
doi:10.1017/S1751731113002188

animal

Predictive ability of selected subsets of single nucleotide polymorphisms (SNPs) in a moderately sized dairy cattle population

J. I. Weller[1†], G. Glick[1,2], A. Shirak[1], E. Ezra[3], E. Seroussi[1], M. Shemesh[1], Y. Zeron[4] and M. Ron[1]

[1]*Department of Ruminant Science, Institute of Animal Sciences, ARO, the Volcani Center, P. O. Box 6, Bet Dagan 50250, Israel;* [2]*Genetics and Breeding, The Robert H. Smith Faculty of Agriculture, The Hebrew University of Jerusalem, Rehovot 76100, Israel;* [3]*Israel Cattle Breeders Association, Caesaria Industrial Park, Caesaria 38900, Israel;* [4]*Sion, AI Institute, Shikmim 79800, Israel*

(Received 27 May 2013; Accepted 30 October 2013)

Several studies have shown that computation of genomic estimated breeding values (GEBV) with accuracies significantly greater than parent average (PA) estimated breeding values (EBVs) requires genotyping of at least several thousand progeny-tested bulls. For all published analyses, GEBV computed from the selected samples of markers have lower or equal accuracy than GEBV derived on the basis of all valid single nucleotide polymorphisms (SNPs). In the current study, we report on four new methods for selection of markers. Milk, fat, protein, somatic cell score, fertility, persistency, herd life and the Israeli selection index were analyzed. The 972 Israeli Holstein bulls genotyped with EBV for milk production traits computed from daughter records in 2012 were assigned into a training set of 844 bulls with progeny test EBV in 2008, and a validation set of 128 young bulls. Numbers of bulls in the two sets varied slightly among the nonproduction traits. In EFF_{12}, SNPs were first selected for each trait based on the effects of each marker on the bulls' 2012 EBV corrected for effective relationships, as determined by the SNP matrix. EFF_{08} was the same as EFF_{12}, except that the SNPs were selected on the basis of the 2008 EBV. In DIF_{max}, the SNPs with the greatest differences in allelic frequency between the bulls in the training and validation sets were selected, whereas in DIF_{min} the SNPs with the smallest differences were selected. For all methods, the numbers of SNPs retained varied over the range of 300 to 6000. For each trait, except fertility, an optimum number of markers between 800 and 5000 was obtained for EFF_{12}, based on the correlation between the GEBV and current EBV of the validation bulls. For all traits, the difference between the correlation of GEBV and current EBV and the correlation of the PA and current EBV was >0.25. EFF_{08} was inferior to EFF_{12}, and was generally no better than PA EBV. DIF_{max} always outperformed DIF_{min} and generally outperformed EFF_{08} and PA. Furthermore, GEBV based on DIF_{max} were generally less biased than PA. It is likely that other methods of SNP selection could improve upon these results.

Keywords: genomic selection, SNP, dairy cattle, genetic evaluation, subsets of SNPs

Implications

Genomic estimated breeding values (GEBV) were derived on the basis of selected subsets of markers from the Illumina BovineSNP50 BeadChip. Single nucleotide polymorphisms (SNPs) were selected based on the effects of each marker on the bulls' genetic evaluations in 2012 and 2008, respectively. The difference between the correlation of GEBV and current EBV and the correlation of the parent average and current EBV was >0.25 for all traits if SNPs were selected based on the 2012 evaluations, but not if SNPs were selected based on 2008 evaluations. Other methods of selection of SNPs may significantly improve genetic evaluations for moderately sized populations.

Introduction

In all of the large commercial dairy cattle populations, thousands of bulls with genetic evaluations based on progeny tests have already been genotyped for the Illumina BovineSNP50 BeadChip. More than 70 000 US Holstein bulls have been genotyped to date (https://www.cdcb.us/Genotype/cur_freq.html). Beginning in 2008, a large number of studies have proposed methods for genomic evaluations in dairy cattle. Most studies have used variations of the method of VanRaden (2008) in which the dependent variable is either

† E-mail: weller@agri.huji.ac.il

the bulls' daughter–yield deviations or deregressed estimated breeding values (EBV), and the independent variables are the genotypes of all valid single nucleotide polymorphisms (SNPs). Genomic estimated breeding values (GEBV) are then derived as an index of the sum of SNP effects, the parent average (PA) EBV and other factors. In nearly all cases, GEBV were evaluated by assigning the population of sires with genotypes and EBV based on progeny tests into a 'training set', consisting generally of older bulls, and a 'validation set' of younger bulls. The effects of the SNP and the regression coefficients for the final index are derived from the training set, and these values are then used to derive GEBV for the validation set, based on PA and genotypes. The GEBV of the validation bulls are then compared with their current deregressed EBV.

Coefficients of determination for the GEBV in the training set are nearly always much higher than coefficients of determination for current EBV in the validation set, especially if bulls are assigned to the two groups on the basis of birth dates. This may be partially because of the higher mean reliabilities of the EBV in the training set. An additional explanation is that linkage relationships and the segregating quantitative trait loci change over time (Moser *et al.*, 2009; Weller *et al.*, 2011). Thus, marker effects derived from the analysis of older bulls may not accurately reflect the marker effects of younger bulls. Glick *et al.* (2012) found that of the 15 485 haplotypes with population frequencies between 5% and 95% in the population of Israeli Holstein bulls born since 1984, 930 haplotypes (6%) underwent significant changes in allelic frequencies, resulting in frequencies of either <10% or >90% for the bulls born between 2004 and 2008.

Various studies have proposed computation of GEBV based on subsets of SNPs. Four basic strategies have been proposed to select SNPs: random selection (Vazquez *et al.*, 2010); equally spaced SNPs throughout the genome (Habier *et al.*, 2009; VanRaden *et al.*, 2009; Weigel *et al.*, 2009; Moser *et al.*, 2010; Vazquez *et al.*, 2010; Zhang *et al.*, 2011); selection of SNPs with the greatest effects on the trait analyzed, as estimated from the analysis of all markers in the training set (Weigel *et al.*, 2009; Moser *et al.*, 2010; Vazquez *et al.*, 2010; Zhang *et al.*, 2011) and selection of markers based on principal component analysis (Pintus *et al.*, 2012). Although accuracies nearly equal to analysis with all markers were obtained with subsets of markers, the accuracy of GEBV computed from subsets of markers was never significantly more than the accuracy of GEBV computed from the analysis of all markers. Daetwyler *et al.* (2008) derived an equation for the expected accuracy of the prediction of the additive genetic value of an individual that can be achieved based on the number of phenotypes recorded and the number of loci affecting the trait of interest.

Unlike the effect of increasing the number of markers, which reaches a plateau for several thousand (VanRaden *et al.*, 2009), increasing the number of bulls analyzed results in more accurate GEBV over the entire range tested to date (VanRaden *et al.*, 2009; Calus, 2010). Thus, the major European countries have formed a consortium that uses training bulls from all participating countries to calculate separate within-country evaluations. Similarly, a North American consortium has been established including the United States, Canada and other countries (Wiggans *et al.*, 2011). Accuracies of GEBV in populations of <1000 genotyped bulls are generally as low as PA derived from traditional evaluations based only on pedigrees and trait phenotypes (VanRaden *et al.*, 2009; Van Grevenhof *et al.*, 2012). Bayesian 'shrinkage' of marker effects improves accuracy of GEBV at best marginally. This is also the case for 'Bayes-B' methodologies, which assume that the majority of markers have no effect on the trait analyzed.

In the current study, we report on four new methods for the selection of markers for inclusion in analysis, and demonstrate that the accuracy of GEBV based on selected sets of markers can be significantly greater than GEBV based on all valid markers. Furthermore, nearly unbiased GEBV can be derived. We also demonstrate that GEBV with higher accuracy than PA can be derived, even though the training set includes <1000 bulls.

Material and methods

The data set and traits analyzed

All valid records from the Israeli Holstein population from January 1985 through May 2012 were included in the analysis to compute EBV. The complete data set was divided into: a 'training set', records generated before June 2008; and the 'validation set', records generated from June 2008. The difference of 4 years between validation set and the complete data set was chosen to mimic the actual dairy situation in that young bulls reach sexual maturity at the age of 1 year, and obtain their first EBV based on daughter records at ~5 years.

Eight traits were analyzed: milk fat, protein production, somatic cell score (SCS), female fertility, persistency of milk production, herd life and PD11, the current Israeli breeding index. EBV were computed for the complete data set, EBV_{12}, and the truncated data set including only records generated before June 2008, EBV_{08}. EBV were derived from multi-trait animal models for milk, fat, protein, SCS, female fertility and persistency, with each parity considered a separate trait, as described by Weller and Ezra (2004) and Weller *et al.* (2006). Parities 1 to 5 were included in the analyses. Female fertility was computed as the inverse of the number of inseminations to conception (Weller and Ezra, 1997). Single-trait animal-model EBV were computed for herd life as described (Settar and Weller, 1999).

Modified daughter-yield-deviations (MDYD), weighted means of daughter records corrected for herd-year-season and parity effects were computed for the bulls of the training set using records generated before June 2008 ($MDYD_{08}$), and for the validation bulls using all records generated up to May 2012 ($MDYD_{12}$). MDYDs were computed according to the following equation:

$$MDYD_i = \sum_{1}^{J} \left(\frac{\sum_{1}^{K} \left(y_{ijkl} - HYS_l - P_k \right)}{K_j + \lambda} \right) / \sum_{1}^{J} \left(K_j + \lambda \right)$$

where $MDYD_i$ is the MDYD for bull i, y_{ijkl} is the record of daughter j of bull i in parity k generated in herd-year-season l, HYS_l is the effect of herd-year-season l, P_k is the effect of parity k, K_j is the number of records (parities) for cow j and λ is the ratio of residual variance to within cow variance. (All variances not included in the cow effects are considered residual variances.) Unlike daughter-yield-deviations computed by VanRaden and Wiggans (1991), MDYDs were not corrected for the genetic effects of the cows' dams. This is because for a large fraction of cows the dam records were not included in the database. Inclusion of dam evaluations with some based on records and some based only on relatives' EBV (including their daughters) could be a potential source of bias.

For milk production traits $\lambda = 1$. For SCS and persistency $\lambda = 4$, and for fertility $\lambda = 10$. Values for λ were derived by the MTC REML program of I. Misztal. The Israeli Holstein population was analyzed for each trait by a repeatability model (e.g. Weller and Ezra, 2004). For herd life, there was only a single record per cow, and therefore no parity effect. $K_j + \lambda$ was replaced with unity. Herd-year-season and parity effects were derived from the standard multi-trait animal model analyses for all traits with multiple parities, and HYS effects for herd life were derived from the standard analysis of this trait.

MDYD for production traits were computed only for bulls with at least 20 effective daughters. For secondary traits, the minimum number of effective daughters was five for SCS and persistency, two for fertility and one valid daughter for herd life. Criteria for acceptance were less restrictive for the secondary traits, because these traits had fewer records for the older bulls.

Animals genotyped and validation of SNPs

A total of 1359 bulls and calves were genotyped: 912 bulls for the 54 001 SNP BeadChip, and 447 for the 54 609 SNP BovineSNP50 v2 BeadChip. The numbers of bulls genotyped by birth year are given in Figure 1. Birth years ranged from 1975 through 2011. The numbers of bulls with genotypes and MDYD in the training and validation data sets and the mean reliabilities of the EBV and median number of daughters by trait are given in Table 1. Reliabilities were estimated by the algorithm of Misztal and Wiggans (1988), as corrected by Misztal *et al.* (1991). Reliabilities for the training bulls are given for the June 2008 evaluation, and for the validation bulls for the May 2012 evaluation. As expected, mean reliabilities were highest for the production traits and lowest for fertility. Mean reliabilities of the validation and training bulls were very similar, but refer to a difference of 4 years. Median numbers of daughters are given instead of mean number of daughters, because some bulls that were returned to general service had thousands of daughters. Of the 128 bulls in the validation set, 31 were half-brothers of sires in the training set. In addition, 110 of the validation bulls were sons of the training bulls.

SNPs were deleted from the analysis if: they did not appear on the original Beadchip, the frequency of the less frequent allele <0.05, there were valid genotypes for less than half of the animals genotyped or if the genotypes of consecutive SNPs were identical for more than 95% of the animals with valid genotypes. For several identical SNPs all were deleted, except for the first. After edits there were 39 302 valid SNPs.

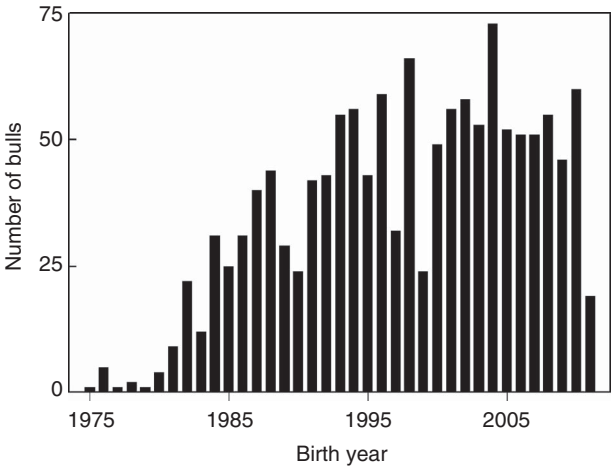

Figure 1 Numbers of bulls with genotypes by birth year.

Table 1 *The number of bulls with genotypes and MDYD in the training and validation data sets, and their mean reliabilities and median number of daughters by trait*

	Number of bulls		Mean reliabilities[1]		Median no. daughters	
Trait analyzed	Training	Validation	Training	Validation	Training	Validation
Milk (kg)	844	128	0.93	0.92	126	103
Fat (kg)	844	128	0.93	0.92	126	103
Protein (kg)	844	128	0.93	0.92	126	103
SCS	785	124	0.90	0.91	120	102
Female fertility (%)	835	144	0.82	0.78	108	88
Persistency (%)	827	131	0.91	0.89	144	107
Herd life (days)	846	131	0.84	0.80	135	104
Israeli index	760	126				

MDYD = modified daughter-yield-deviations; SCS = somatic cell score.
[1]Reliabilities and numbers of daughters for the training bulls are given for the June 2008 evaluation, and for the validation bulls for the May 2012 evaluation.

Calculations of genomic evaluations and selection of SNPs
The method of VanRaden (2008) was used to compute genomic effects on the MDYD from the training set for each trait. The model used was as follows:

$$\mathbf{y} = \mu + \mathbf{Za} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of $MDYD_{08}$, μ is the mean, $\mathbf{Z}$ is the incidence matrix that relates $\mathbf{y}$ with the genomic effects in the vector $\mathbf{a}$ and $\mathbf{e}$ is the random residual term. The relationship matrix based on marker information, $\mathbf{G}$, was constructed as in VanRaden (2008):

$$\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})'}{2\sum_{j=1}^{n} p_j(1-p_j)}$$

where $\mathbf{M}$ is a matrix of SNP genotypes for each animal, and $\mathbf{P}$ is a matrix of two times the difference between the frequency of the second allele p at locus j and 0.5. Dimensions of $\mathbf{M}$ and $\mathbf{P}$ are the number of individuals by the number of markers. Elements of $\mathbf{M}$ are set to −1, 0 and 1 for the homozygote, heterozygote and other homozygote. For missing genotypes the element of $\mathbf{M}$ was assumed to be equal to the corresponding element of $\mathbf{P}$. As in VanRaden (2008), allelic frequencies for each SNP were computed from the sample of 'founder' bulls in the training population. That is, bulls without genotyped ancestors. The diagonals of G were augmented by $2\sum p_i(1-p_i)$, under the prior assumptions that all markers included in the analysis account for all the variances among MDYD, and the effects of all markers are equal. Direct genomic evaluations (DGE) were then computed as: $\mathbf{Z}\hat{\mathbf{a}}$, where $\hat{\mathbf{a}}$ is the vector of solutions for $\mathbf{a}$.

Similar to VanRaden *et al.* (2009), final GEBV were computed from an index including the direct genomic effects and PA. The regression coefficients for the index were derived from the training data set, using the following equation:

$$EBV_{12} = \text{int} + a \times \text{DGE} + b \times \text{PA} + e$$

where int is the y-intercept, a and b are regression constants, and the other terms are as defined previously. DGE and PA were computed from the truncated data set including only records generated before June 2008. PA were computed as the means of the parent EBV derived from the standard multi-trait analysis of the population. All bulls in the training set with genotypes, $MDYD_{08}$ and EBV_{08}, for dams based on at least one lactation record were included in the analysis.

The regression coefficients derived from the training set were then used to compute GEBV for the validation bulls by the same equation, with PA for the validation bulls also computed based only on records generated before June 2008. The GEBV of the validation bulls and their genetic merit estimates based on the PA were compared with their EBV_{12} and $MDYD_{12}$. Correlations of the GEBV of the validation bulls with their EBV_{12} were compared with the correlations of their PA with their EBV_{12}. As the PA has a major effect on EBV of low heritability traits, even with more than 50 daughters, correlations of GEBV and PA with $MDYD_{12}$ were also computed. In addition, to estimate the bias of PA and GEBV, relative to the EBV_{12}, regressions of PA and GEBV on EBV_{12} were computed, and means and standard deviations of PA, GEBV and EBV_{12} were compared.

Four methods were used to select subsets of SNPs for analysis. These methods were applied to all eight traits listed in Table 1. In the first method, 'EFF_{12}', SNPs were selected for each trait based on the fixed additive effect of each marker on the bulls' EBV_{12} for each trait, as derived by the analysis of all valid SNPs by the 'EMMAX' algorithm (Kang *et al.*, 2010). The additive effect was computed as the regression of the bulls' EBV_{12} on the number of '+' alleles of the SNP. Determination of the '+' allele was arbitrary. This algorithm corrects for relationships among animals by calculating an empirical relationship matrix based on SNP genotypes. The SNPs with the greatest absolute effects were retained for the computation of GEBV. It should be noted that this method uses information available only after May 2008 to select the SNPs included for analysis. However, once the sample of SNPs is selected, GEBV are computed as described above based only on records generated before June 2008. The second method, 'EFF_{08}', was the same as EFF_{12}, except that the dependent variables in the 'EMMAX' analysis were the bulls' EBV_{08}. Thus, only records available in May 2008 were used to select the SNPs and to compute GEBV.

In the third method, 'DIF_{max}', SNPs were selected in two stages. In the first stage, SNPs with minor allele frequencies >0.05 and at least 200 valid genotypes among the training bulls population were retained. Of the 39 302 valid SNPs, 30 288 met these additional criteria. In the second stage, these SNPs were ranked by the absolute difference in allelic frequencies between the training and validation bulls from largest to smallest. In the preliminary analysis, the 1000 SNPs with the greatest absolute differences were selected for inclusion. The number of SNPs included was then increased by increments of 500 up to 2000, or until a decrease of >2% in the correlation of the GEBV of the validation bulls with their EBV_{12} was obtained. In each additional run, the 500 SNPs with the next greatest allelic differences were added to the previous sample of SNPs. If a decrease of 2% in the correlation was not obtained with 2000 SNPs, then the number of SNPs included was increased by increments of 1000 up to 6000. If a 2% reduction in the correlation was obtained with 1500 SNPs, relative to 1000 SNPs, then the number of SNPs included was decreased by increments of 100 until a 2% decrease in the correlation was obtained. Again, the SNPs with the greatest allelic differences among those included in the previous run were retained.

As a negative control, DGE and GEBV were also computed with the SNPs with the smallest absolute difference in allelic frequencies between the training and validation bulls among the SNPs with minor allele frequencies >0.05 and at least 200 valid genotypes among the training bulls. This method was denoted 'DIF_{min}'. For each trait, the number of SNPs included in DIF_{min} was the number that resulted in the highest correlation between the GEBV of the validation bulls with their EBV_{12} for DIF_{max}.

In comparison with these four methods, GEBV were also computed for all traits using all 39 816 valid markers and using each 20th valid SNP (1991 SNPs).

Results and discussion

Correlations of GEBV and PA for the validation bulls with EBV_{12} and $MDYD_{12}$ from the analysis of all SNPs and equally spaced SNPs are given in Table 2. With all valid SNPs, correlations of GEBV with EBV_{12} and $MDYD_{12}$ were slightly higher than PA for fat, fertility, persistency and PD11, but lower for the other traits. Differences between GEBV computed with all SNPs and each 20th SNP were minimal, except for fertility and persistency, for which analysis with all SNPs was superior.

The correlations between the GEBV-based EFF_{12} and EBV_{12} as a function of the number of SNPs included in the analysis for the validation bulls are plotted in Figure 2. There was a clear optimum for all the traits, except for herd life. The optimum number of markers was between 600 and 6000 for all of the traits analyzed.

Correlations of EFF_{12} GEBV and PA with EBV_{12} and $MDYD_{12}$ with optimum number of SNPs, and coefficients applied to derive GEBV from PA and DGE, are given in Table 3. The intercepts were negative for all traits, except for SCS. Coefficients for DGE were greater than coefficients for PA for all traits, but coefficients for PA were still significant ($P < 0.05$) for all traits, except for protein and SCS. The correlations of GEBV with EBV_{12} and $MDYD_{12}$ were higher than the correlations of PA with EBV and $MDYD_{12}$ for all traits, and ranged from 0.75 for herd life to 0.92 for persistency. Correlations of this magnitude are generally obtained only for much larger populations (e.g. VanRaden *et al.*, 2009). The mean difference in the correlations between GEBV and PA was 0.36. Differences in correlations between PA and GEBV with EBV_{12} and $MDYD_{12}$ were similar for all traits. The greatest differences in correlations were obtained for PD11 for both EBV_{12} and $MDYD_{12}$, close to 0.45.

According to quantitative genetic theory, the PA should explain no more than 50% of the genetic variance in the progeny, for a maximum correlation of 0.71 (Lynch and Walsh, 1998). However, the EBV_{12} for the validation bulls are based on relatively low numbers of records. In this case, the parent contribution to the EBV is significant, especially for low heritability traits. Therefore, correlations between GEBV and EBV_{12} were higher for fertility and persistency, which have low heritability, whereas correlations of GEBV and PA with $MDYD_{12}$ were lower. All correlations were lower for herd life, which has only one record per cow. For protein, the correlations of EBV_{12} with GEBV and PA were 0.82 and 0.42. Although selection of SNPs was based on information not available in May 2008, the effects derived from these subsets of SNPs were based entirely on information available in 2008.

In EFF_{12}, the SNPs were selected based on their 2012 EBV. In EFF_{08} SNPs were selected by the same procedure, but the dependent variables were the 2008 EBV. Thus, this method only used information available in May 2008. Correlations of the GEBV-based EFF_{08} with EBV_{12} and $MDYD_{12}$, and coefficients applied to derive GEBV from PA and DGE, are given in

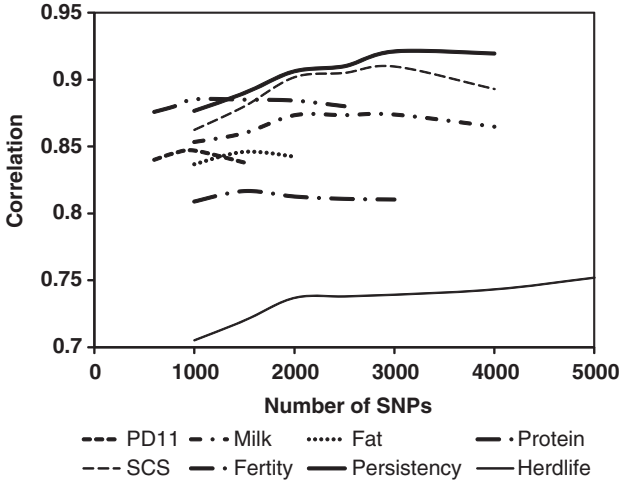

Figure 2 Correlations between genomic estimated breeding values derived by EFF_{12} and June 2012 estimated breeding values for the validation bulls as a function of the numbers of SNPs included in the analysis.

Table 2 *Correlations of GEBV and PA with June 2012 EBV_{12} and MDYD from the analysis of all SNPs (All) and each 20th SNP (E20) for the validation bulls*

	Correlations					
	PA		All SNPs		E20 SNPs	
Trait	EBV_{12}	MDYD	EBV_{12}	MDYD	EBV_{12}	MDYD
Milk	0.55	0.50	0.47	0.45	0.46	0.43
Fat	0.44	0.35	0.49	0.40	0.48	0.38
Protein	0.39	0.41	0.36	0.36	0.36	0.36
SCS	0.54	0.42	0.48	0.40	0.45	0.38
Fertility	0.66	0.36	0.67	0.41	0.62	0.37
Persistency	0.60	0.45	0.65	0.50	0.56	0.40
Herd life	0.38	0.36	0.38	0.31	0.35	0.28
PD11	0.39	0.34	0.42	0.40	0.42	0.40

GEBV = genomic estimated breeding values; PA = parent averages; EBV = estimated breeding values; MDYD = modified daughter-yield-deviations; SNP = single nucleotide polymorphism; SCS = somatic cell score.

Table 3 *Correlations of GEBV derived by EFF_{12} and MDYD with optimum number of SNPs for the validation bulls, and coefficients applied to derive GEBV from PA and DGE*

		Correlations		Coefficients		
Traits	Optimum No. SNPs	EBV_{12}	MDYD	Intercept	PA	DGE
Milk	3000	0.87	0.87	−192.3	0.16	0.79
Fat	1500	0.85	0.80	−5.3	0.13	0.75
Protein	1500	0.82	0.79	−4.9	−0.02[1]	0.66
SCS	3000	0.91	0.87	0.09	0.03[1]	1.12
Fertility	2000	0.89	0.67	−0.43	0.44	0.61
Persistency	3000	0.92	0.81	−1.13	0.14	1.04
Herd life	5000	0.75	0.58	−48.8	0.49	0.56
PD11	800	0.85	0.81	−214.7	0.22	0.61

GEBV = genomic estimated breeding value; MDYD = modified daughter-yield-deviation; SNP = single nucleotide polymorphism; PA = parent averages; DGE = direct genomic evaluations; EBV = estimated breeding value; SCS = somatic cell score.
[1]Not significant at $P < 0.05$. All other coefficients were significant.

Table 4 *Correlations of GEBV derived by EFF_{08} with June 2012 EBV_{12} and MDYD with optimum number of SNPs for the validation bulls, and coefficients applied to derive GEBV from PA and DGE*

Traits	Optimum No. SNPs	Correlations		Coefficients		
		EBV_{12}	MDYD	Intercept	PA	DGE
Milk	2000	0.55	0.54	−187.8	0.15	0.78
Fat	1500	0.49	0.43	−5.4	0.06	0.77
Protein	2000	0.40	0.37	−3.8	0.15	0.57
SCS	3000	0.50	0.42	0.09	0.01[1]	1.20
Fertility	4000	0.66	0.42	−0.47	0.37	0.66
Persistency	300	0.66	0.55	−1.06	0.20	0.92
Herd life	2000	0.43	0.38	−29.7	0.73	0.37
PD11	1000	0.40	0.37	−219.8	0.19	0.62

GEBV = genomic estimated breeding value; EBV = estimated breeding value; MDYD = modified daughter-yield-deviation; SNP = single nucleotide polymorphism; PA = parent averages; DGE = direct genomic evaluations; SCS = somatic cell score.
[1]Not significant at $P < 0.05$. All other coefficients were significant.

Table 5 *Coefficients applied to derive GEBV from PA and DGE for DIF_{max} and DIF_{min}*

Traits	Coefficients					
	DIF_{max}[1]			DIF_{mim}[2]		
	Intercept	PA	DGE	Intercept	PA	DGE
Milk	−133.4	0.72	0.44	−155.7	0.64	0.60
Fat	−4.3	0.46	0.56	−4.9	0.47	0.64
Protein	−2.8	0.56	0.33	−3.1	0.67	0.32
SCS	0.09	0.23	1.23	0.09	0.17	1.30
Fertility	−0.29	0.63	0.65	−0.30	0.59	0.69
Persistency	−1.01	0.32	1.11	−1.06	0.30	1.20
Herd life	−17.0	0.89	0.22	−16.1	0.94	0.22
PD11	−173.4	0.42	0.50	−174.6	0.51	0.50

GEBV = genomic estimated breeding value; PA = parent averages; DGE = direct genomic evaluations; SCS = somatic cell score.
[1]Single nucleotide polymorphisms (SNPs) selected by maximum difference in allelic frequencies between training and validation bulls.
[2]SNPs selected by minimum difference in allelic frequencies between training and validation bulls.

Table 4. Similar to EFF_{12}, the intercepts were negative for all traits, except for SCS, and coefficients for DGE were greater than coefficients for PA for all traits, except for herd life. Coefficients for PA were significant ($P < 0.05$) for all traits, except for SCS. The EFF_{08} correlations were significantly lower than the EFF_{12} correlations for all traits for both EBV_{12} and $MDYD_{12}$. Correlations of EFF_{08} GEBV with EBV_{12} were similar to the PA correlations for all traits. Correlations of GEBV with $MDYD_{12}$ were higher or equal to the correlations of PA with $MDYD_{12}$ for all traits, but by relatively small margins. In the analysis of the validation set, the PA values were highly correlated with the sum of the SNP effects. Thus, both the genetic marker effects and PA were largely determined by the same QTL segregating in the population of training bulls. However, as noted previously (Glick *et al.*, 2012), the QTL segregating at

Table 6 *Correlations of GEBV derived by DIF_{max} and DIF_{min} with June 2012 EBV_{12} and MDYD with optimum number of SNPs for the validation bulls*

Traits	Optimum No. SNPs	Correlations			
		DIF_{max}[1]		DIF_{mim}[2]	
		EBV_{12}	MDYD	EBV_{12}	MDYD
Milk	800	0.55	0.53	0.53	0.50
Fat	1500	0.53	0.41	0.44	0.33
Protein	800	0.47	0.42	0.43	0.39
SCS	3000	0.51	0.43	0.37	0.30
Fertility	4000	0.67	0.41	0.66	0.40
Persistency	3000	0.64	0.49	0.52	0.38
Herd life	500	0.47	0.41	0.35	0.32
PD11	3000	0.51	0.47	0.31	0.30

GEBV = genomic estimated breeding value; EBV = estimated breeding value; MDYD = modified daughter-yield-deviation; SNP = single nucleotide polymorphism; SCS = somatic cell score.
[1]SNPs selected by maximum difference in allelic frequencies between training and validation bulls.
[2]SNPs selected by minimum difference in allelic frequencies between training and validation bulls.

intermediate frequencies in the validation bulls are not the same as those segregating in the training population. Thus, we conclude that to improve GEBV, it is necessary to include markers linked to QTL that are not segregating at intermediate frequencies in the training set.

The mean difference in the SNP allelic frequencies was 0.043, and the maximum difference was 0.27. Five percent of the SNPs (1514) had differences >0.11. Coefficients for factors included in the models used to compute GEBV for DIF_{max} and DIF_{min} are given in Table 5, and correlations of GEBV computed by DIF_{max} and DIF_{min} with EBV_{12} and $MDYD_{12}$ with optimum number of SNPs for DIF_{max} are given in Table 6. Generally, the coefficients were similar for DIF_{max} and DIF_{min}. Intercepts were similar to those for EFF_{12} and EFF_{08}, but coefficients for PA were generally larger than the corresponding coefficients for EFF_{12} and EFF_{08}, and larger than the DIF_{max} and DIF_{min} coefficients for DGE for milk, protein and herd life.

The correlations of the GEBV with EBV_{12} were higher with DIF_{max} than with the analysis of all SNPs for all traits, except for fertility and persistency. Furthermore, the correlations of the GEBV with EBV_{12} were higher with DIF_{max} than EFF_{08} for all traits, except for milk production and persistency. Correlations for both GEBV and $MDYD_{12}$ for PD11 were ~0.1 greater for DIF_{max}, as compared with EFF_{08}. Correlations for DIF_{max} were greater than correlations for DIF_{min} for all traits, although differences were minimal for milk and fertility. DIF_{max} is expected to be more effective for traits that underwent strong recent selection, such as the breeding index and protein production. Milk is not included in PD11, and genetic trend for fertility in Israel over the last decade has been minimal (Glick *et al.*, 2012). Both DIF_{max} and DIF_{min} used only information available in June 2008. Of course, differences in allelic frequencies are not necessarily the result

Table 7 *Regressions and coefficients of the determination of PA and GEBV derived by DIF_{max} on June 2012 EBV_{12} for the validation bulls*

	Regression on EBV_{12}		Coefficient of determination	
Traits	PA	GEBV	PA	GEBV
Milk	0.98	1.00	0.28	0.31
Fat	0.88	1.06	0.20	0.28
Protein	0.86	1.00	0.15	0.22
Somatic cell score	0.87	0.79	0.29	0.26
Fertility	1.07	1.02	0.41	0.45
Persistency	1.02	1.11	0.36	0.41
Herd life	0.79	0.93	0.15	0.22
PD11	0.93	1.15	0.15	0.26

PA = parent averages; GEBV = genomic estimated breeding value; EBV = estimated breeding value.

Table 8 *Means and standard deviations of PA, GEBV derived by DIF_{max} and June 2012 EBV_{12} for the validation bulls*

	Means			s.d.		
Traits	PA	GEBV	EBV_{12}	PA	GEBV	EBV_{12}
Milk	237	−0	120	183	205	336
Fat	15.8	13.9	13.0	6.9	6.1	13.3
Protein	12.6	10.7	10.9	4.0	3.9	8.7
SCS	−0.081	−0.092	−0.101	0.12	0.15	0.20
Fertility	0.36	0.28	0.57	1.50	1.74	2.44
Persistency	0.57	0.54	−0.16	1.34	1.26	2.16
Herd life	55	60	51	41.3	42.2	83.5
PD11	466	377	370	127	124	335

PA = parent averages; GEBV = genomic estimated breeding value; EBV = estimated breeding value; SCS = somatic cell score.

of selection, and may be caused by random drift, the 'hitchhiker effect' or other factors.

Genetic evaluations are unbiased if the means are equal to the means of the true genetic values and the regressions of EBV on true genetic values are equal to unity. As true genetic values are unknown, GEBV and PA were compared with EBV_{12}. Regressions and coefficients of determination of PA and DIF_{max} GEBV on EBV_{12} are presented in Table 7, and means and standard deviations of PA, DIF_{max} GEBV and EBV_{12} are given in Table 8. Standard errors of the regressions were ~0.1. GEBV regressions were generally higher than the PA regressions. Exceptions were SCS and fertility, but the regressions for fertility were very close to unity for both methods. GEBV regressions were close to unity for all traits, except for SCS. Thus, by this criterion, the GEBV can be considered virtually unbiased for all traits, except for SCS. With respect to means, GEBV were less biased than PA for all traits, except for fertility and herd life. Coefficients of determination for GEBV were higher than for PA for all traits, except for SCS.

Although numerous studies have computed GEBV derived from the subsets of markers (Habier *et al.*, 2009; VanRaden *et al.*, 2009; Weigel *et al.*, 2009; Moser *et al.*, 2010; Vazquez *et al.*, 2010; Zhang *et al.*, 2011), this is the first study to show that more accurate GEBV can be derived from the analysis of subsets of markers, as compared with the analysis of all valid markers. This is not too surprising considering that the methodologies used to select markers in this study differed from all previous studies. In the previous studies that selected markers based on their effects, the effects were derived from the genomic analysis of all markers, with marker effects regressed in proportion to the fraction of genetic variance assumed to be associated with each marker. This results in strongly underestimated effects relative to the actual QTL effects.

Table 9 *Comparison of correlations of GEBV derived by all methods with June 2012 EBV_{12} and MDYD for PD11*

	Correlations	
Method	EBV_{12}	MDYD
Parent averages	0.39	0.34
All SNPs	0.42	0.40
Each 20th SNP	0.42	0.40
EFF_{12}	0.85	0.81
EFF_{08}	0.40	0.37
DIF_{max}	0.51	0.47
DIF_{min}	0.31	0.30

GEBV = genomic estimated breeding value; EBV = estimated breeding value; MDYD = modified daughter-yield-deviation; SNP = single nucleotide polymorphism.

The correlations of GEBV derived by all methods with EBV_{12} and MDYD for PD11 are summarized in Table 9. Although EFF_{12} is clearly superior to all other methods, DIF_{max}, which includes only information available in 2008, is clearly the next best. As noted first by Moser *et al.* (2009), a major weakness in the application of genomic evaluation is that different genes are likely to be segregating in old and young bulls. The four methods applied in this study shed additional light on this problem. EFF_{12}, which selects markers based on their effects in both young and old bulls, clearly gives the best results, but in this method markers are selected on the basis of data not available in real time. EFF_{08}, which selects markers based on their effects in the training bulls, is at best only marginally better than PA. This indicates that the QTL segregating in the older bulls are generally no longer segregating in the younger bulls. The rational for DIF_{max}, selection of a subset of SNPs based on difference in allelic frequencies between old and young bulls, originated from the finding of Glick *et al.* (2012) that allelic frequencies of markers in linkage disequilibrium to QTL under selection can change significantly over a single generation. This is also true for the causative mutation in *ABCG2* underlying the QTL for production on BTA6 that showed a difference in the favorable allele frequency of 0.2 over two generations of selection (Cohen-Zinder *et al.*, 2005). Thus, these markers can be considered prime candidates for inclusion in genomic evaluation. Although EFF_{12} cannot be directly applied to actual data, the results of this method raise the possibility that other methods of selection of SNPs

could significantly improve GEBV derived for moderately sized populations.

One of the main reasons that GEBV are inaccurate for small populations is that most markers have no actual effects on the trait analyzed. However, the estimated effects, which consist nearly entirely of random error, overwhelm the effects associated with segregating QTL. This problem decreases as sample sizes increase, and prediction error variances decrease. Therefore, it is not too surprising that relatively high-accuracy GEBV can be derived for a population of the size of the Israeli Holsteins, if only markers with real effects are included in the analysis.

Daetwyler *et al.* (2008) derived the following equation for the expected accuracy of the prediction of the additive genetic value ($r_{\hat{g}g}$) of an individual that can be achieved based on the measurement of n_P phenotypes, assuming that n_G potential loci affect the trait of interest:

$$r_{\hat{g}g} = \sqrt{\frac{\lambda h^2}{\lambda h^2 + 1}}$$

where h^2 is the observed heritability and $\lambda = n_P/n_G$. In the current study, as the 'phenotypes' were genetic evaluations, the heritability is the mean reliability of the evaluations, which is close to 0.9. The actual number of loci effecting the trait is of course unknown, but for the accuracy of prediction to equal 0.75, requires $\lambda = n_P/n_G = 1.43$. That is, the number of phenotypes should be approximately equal to 1.4 times the number of loci affecting the trait. As only about 800 individuals were included in the training population, n_G should be approximately equal to 560. This number is lower than the optimal number of SNPs for most of the traits analyzed, but it is likely that in all cases SNPs were included that did not have effects on the traits analyzed.

Conclusions

GEBV derived from selected sets of markers can outperform GEBV derived from the analysis of all markers. GEBV derived from selected sets of markers can outperform PA, even if the training population includes <1000 bulls. Using the strategy that resulted in the greatest accuracy of evaluation, correlations of GEBV with EBV_{12} were 0.35 higher than correlations of PA with EBV_{12}. However, this method uses information that is not available in real time. Even if the selection of markers is based only on information available at the time the training set is generated, it is still possible to select sets of markers that yield higher correlations between GEBV and EBV_{12} than correlations of PA with EBV_{12}. Furthermore, GEBV were less biased than PA. This study raises the possibility that other methods of selection of SNPs could significantly improve GEBV derived for moderately sized populations.

Acknowledgments

This research was supported by grants from the Israel Dairy Board, the Chief Scientist of the Israeli Ministry of Agriculture and Rural Development and Binational Agricultural Research and Development Fund (BARD) Research Project IS-4394-11R. Genotyping was performed by A. Schein and N. Avidan, Pharmacogenetics and Translation Medicine Center, The Rappaport Institute for Research in the Medical Sciences, Technion, Haifa, Israel, and GeneSeek, Lincoln, NE, USA. The authors thank the reviewers for their suggestions, and I. Misztal for use of the MTC program.

References

Calus MPL 2010. Genomic breeding value prediction: methods and procedures. Animal 4, 157–164.

Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee JH, Drackley JK, Band MR, Hernandez AG, Shani M, Lewin HA, Weller JI and Ron M 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Research 15, 936–944.

Daetwyler HD, Villanueva B and Woolliams JA 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. Plos One 3, e3395.

Glick G, Shirak A, Uliel S, Zeron Y, Ezra E, Seroussi E, Ron M and Weller JI 2012. Signatures of contemporary selection in the Israeli Holstein dairy cattle. Animal Genetics 43 (Suppl. 1), 45–55.

Habier D, Fernando RL and Dekkers JCM 2009. Genomic selection using low-density marker panels. Genetics 182, 343–353.

Kang HM, Sul JJ, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C and Eskin E 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42, 348–354.

Lynch M and Walsh B 1998. Genetics and analysis of quantitative traits. Sinauer Associates Inc., Sunderland, MA.

Misztal I and Wiggans GR 1988. Approximation of prediction error variance in large-scale animal models. Journal of Dairy Science 71 (Suppl. 2), 27–32.

Misztal I, Lawlor TJ, Short TH and Wiggans GR 1991. Continuous genetic evaluation of Holstein for type. Journal of Dairy Science 74, 2001–2009.

Moser G, Khatkar MS, Hayes BJ and Raadsma HW 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genetics Selection Evolution 42, 37.

Moser G, Tier B, Crump RE, Khatkar MS and Raadsma HW 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution 41, 56.

Pintus MA, Gaspa G, Nicolazzi EL, Vicario D, Rossoni A, Ajmone-Marsan P, Nardone A, Dimauro C and Macciotta NPP 2012. Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach. Journal of Dairy Science 95, 3390–3400.

Settar P and Weller JI 1999. Genetic analysis of cow survival in the Israeli dairy cattle population. Journal of Dairy Science 82, 2170–2177.

Van Grevenhof EM, Van Arendonk JAM and Bijma P 2012. Response to genomic selection: the Bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. Genetics Selection Evolution 44, 26.

VanRaden PM 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423.

VanRaden PM and Wiggans GR 1991. Derivation, calculation and use of national animal model information. Journal of Dairy Science 74, 2737–2746.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF and Schenkel FS 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92, 16–24.

Vazquez AI, Rosa GJM, Weigel KA, de los Campos G, Gianola D and Allison DB 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. Journal of Dairy Science 93, 5942–5949.

Weigel KA, de los Campos G, Gonzalez-Recio O, Naya H, Wu XL, Long N, Rosa GJ and Gianola D 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of Dairy Science 92, 5248–5257.

Weller JI and Ezra E 1997. Genetic analysis of somatic cell concentration and female fertility of Israeli Holsteins by the individual animal model. Journal of Dairy Science 80, 586–593.

Weller JI and Ezra E 2004. Genetic analysis of the Israeli Holstein dairy cattle population for production and nonproduction traits with a multitrait animal model. Journal of Dairy Science 87, 1519–1527.

Weller JI, Ezra E and Leitner G 2006. Genetic analysis of persistency in the Israeli Holstein population by the multitrait animal model. Journal of Dairy Science 89, 2738–2746.

Weller JI, Ron M, Glick G, Shirak A, Zeron Y, Misztal I and Ezra E 2011. A simple algorithm for genomic selection for moderately sized dairy cattle populations. Animal 6, 193–202.

Wiggans GR, VanRaden PM and Cooper TA 2011. The genomic evaluation system in the United States: past, present, future. Journal of Dairy Science 94, 3202–3211.

Zhang Z, Ding X, Liu J, Zhang Q and de Koning D-J 2011. Accuracy of genomic prediction using low-density marker panels. Journal of Dairy Science 94, 3642–3650.