Predictive ability of selected subsets of single nucleotide polymorphisms (SNPs) for moderately sized dairy cattle populations

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

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

Abstract

Several studies have shown that computation of genomic estimated breeding values (GEBV) with accuracies significantly greater than parent average EBV requires genotyping of at least several thousand progeny-tested bulls. For all published analyses, GEBV computed from selected samples of markers have lower or equal accuracy than GEBV derived based on all valid 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 969 Israel Holstein bulls genotyped with EBV for milk production traits computed from daughter records in November, 2011, were assigned into a training set of 829 bulls with progeny test EBV in June, 2008, and a validation set of 140 young bulls. Numbers of bulls in the two sets varied slightly among the nonproduction traits. In Method 1, SNPs were first selected for each trait based on a linear model analysis of the effect of each marker on the bulls' current EBV for each trait. A subset of these SNP then was analyzed by a REML model including relationships. Method 2 was the same as Method 1, except that that the dependent variable was the 2008 EBV. In Method 3, the SNPs with the greatest effects on the 2008 EBV, as determined by the REML analysis were deleted. Of the remaining SNPs, the markers with the greatest effects on 2011 EBV were retained. In Method 4, the SNPs with the greatest effects on the 2008 EBV, as determined by the REML analysis were deleted. Of the remaining SNPs, the markers with the greatest change in allele frequency between the bulls in the training set, and the validation bulls were retained for analysis. For all methods, the numbers of SNPs deleted and retained were varied to obtain a maximum correlation between the GEBV and EBV of the validation bulls. In Methods 1 and 2, the number of SNPs included in the analyses was varied over the range of 400 to 6000. For each trait, except fertility, an optimum number of markers between 600 and 2000 was obtained for Method 1, 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 parent average and current EBV was >0.1. Method 2 was inferior to Method 1 and generally no better than parent average EBV, but Method 3 outperformed Method 1. Even Method 4, in which selection of markers is based only on information available at the time the training set is generated, correlations between GEBV and current EBV were on the average 0.042 higher than correlations of parent averages with current EBV. Furthermore, GEBV were less biased than parent averages. It is likely that other methods of SNP selection could improve upon these results.

Keywords: Genomic selection, SNP, Dairy cattle, Genetic evaluation

Introduction

All of the large dairy cattle populations have already genotyped thousands of bulls with genetic evaluations based on progeny tests for the Illumina BovineSNP50 BeadChip. 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 the bulls' daughter-yield-deviations or deregressed estimated breeding values (EBV), and the independent variables are the genotypes of all valid SNPs. Genomic EBV (GEBV) are then derived as an index of the sum of SNP effects, the parent average EBV (PA) and other factors. In nearly all cases, GEBV were evaluated by dividing 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 only on PA and genotypes. The GEBV of the validation bulls are then compared to their current EBV.

Coefficients of determination for the GEBV in the training set are nearly always much higher than correlations of GEBV with current EBV in the validation set, especially if bulls are assigned to the two groups based on birth dates. A possible explanation is that linkage relationships and the segregating quantitative trait loci change over time (Moser et al., 2009). Glick et al. (2012) found that out of the15,485 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. Three basic strategies have been proposed to select SNPs:

1. Random (Vazquez et al., 2010).
2. Equally spaced throughout the genome (Habier et al., 2009; Moser et al., 2010; VanRaden et al., 2009; Vazquez et al., 2010; Weigel et al., 2009; Zhang et al., 2011).
3. Markers with the greatest effects on the trait analyzed, as estimated from the analysis of all markers (Moser et al., 2010; Vazquez et al., 2010; Weigle et al., 2009; Zhang et al., 2011).

Although accuracies nearly equal to analysis with all markers were obtained with subsets of markers, the accuracy of GEBV computed from subsets of markers is never significantly more than the accuracy of GEBV computed from analysis of all markers.

Unlike the effect of increasing the number of markers, which reaches a plateau for several thousand, increasing the number of bulls analyzed results in more accurate GEBV over the entire range tested to date. Furthermore, the accuracy of GEBV for young bulls computed from analysis of <1000 bulls in the training set is no higher than the accuracy of PA (e. g., VanRaden et al., 2009). Bayesian "shrinkage" of marker effects improves accuracy of GEBV at best marginally.

In the current study we report on a four new methods for selection of markers for inclusion in analysis, and demonstrate that the accuracy of GEBV based on selected sets of marker can be greater than GEBV based on all valid markers. 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 November, 2011, were included in the analysis. Eight traits were analyzed; milk fat, and protein production, somatic cell score (SCS), female fertility, persistency of milk production, herd-life, and PD11, the current Israeli breeding index. Multitrait animal model EBV were computed for milk, fat, protein, SCS, female fertility, and persistency, with each parity considered a separate trait as described by Weller & Ezra (2004) and Weller et al. (2006). Parities 1-5 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).

The complete data set was divided into a "training set," records generated prior to June, 2008; and the "validation set," records generated from June, 2008. The difference of 3.5 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 one year, and obtain their first EBV based on daughter records at approximately 4.5 years.

Modified daughter-yield-deviations (MDYD), weighted means of daughter records corrected for herd-year-season and parity effects; were computed for each bull with valid daughter records in the training set, using records prior to June, 2008. Only bulls with at least 20 effective daughters for milk production traits, 5 effective daughters for SCS and persistency, 2 effective daughters for fertility, and 1 valid daughter for herd-life were included in the analysis of each trait. In addition, current MDYD were computed based on all records in the complete data set.

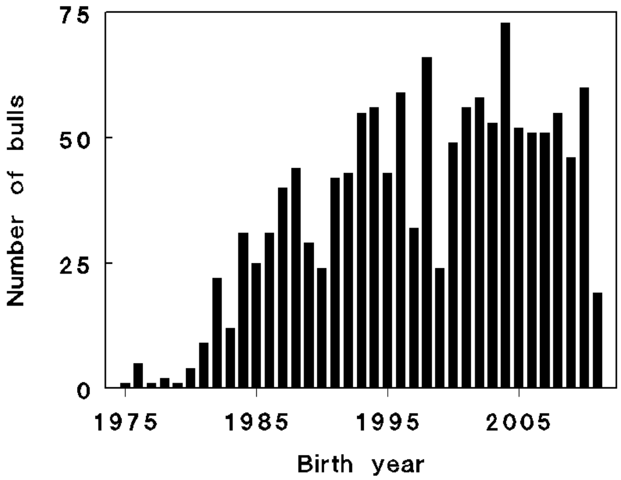


Figure 1. Numbers of bulls with genotypes by birth year.

Animals genotyped and validation of SNPs

A total of 1359 bulls and calves were genotyped, 912 bulls for the 54001 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 by trait are given in Table 1.

Table 1. The number of bulls with genotypes and MDYD in the training and validation data sets by trait.

Trait analyzed	Number of bulls	
	Training	Validation
Milk (kgs)	829	140
Fat (kgs)	829	140
Protein (kgs)	829	140
SCS	785	121
Female fertility (%)	835	139
Persistency (%)	827	129
Herdlife (days)	846	129
Israeli Index	760	110

SNPs were deleted from analysis if:

1. They did not appear on the original Beadchip.
2. The frequency of the less frequent allele < 0.05.
3. There were valid genotypes for < half of the animals genotyped.
4. The genotypes of two consecutive SNPs were identical for > 95% of the animals with valid genotypes. In this case the second SNP was deleted.

After edits there were 39,816 valid SNPs.

Calculations of genomic evaluations and selection of SNPs

The method of VanRaden (2008) was used to compute marker effects on the MDYD from the training set for each trait. Regression coefficients for the sum of marker effects, PA and birth year effects were then computed from the training data set, using all bulls with genotypes, MDYD, and EBV for dams based on at least one lactation record.

The regression coefficients derived from the training set were then used to compute GEBV for the "validation bulls," bulls with MDYD in the total population and dam EBV based on at least one lactation record, but without EBV based on daughter records in the validation set. The GEBV of the validation bulls and their PA were compared to their November, 2011, EBV and MDYD computed from the complete data set. GEBV were computed as described for protein and female fertility using all valid markers and using each 20th valid SNP.

Four additional methods were used to select subsets of SNPs for analysis. In Method 1, SNPs were first selected for each trait based on a linear model analysis of the effect of each marker on the bulls' November, 2011, EBV for each trait. A subset of these SNP was analyzed by a REML model including relationships. In both steps each SNP was analyzed separately. The number of SNPs included in the analysis was varied over the range of 400 to 6000 to obtain an optimum. Method 2 was the same as Method 1, except that that the dependent variables were the bulls' June, 2008, EBV.

In Method 3, the SNPs with the greatest effects on the 2008 EBV, as determined by the REML analysis were deleted. Of the remaining SNPs, the markers with the greatest effects on 2011 EBV were retained. In Method 4, the SNPs with the greatest effects on the 2008 EBV, as determined by the REML analysis were deleted. Of the remaining SNPs, the markers with the greatest change in allele frequency between the bulls in the training set, and the validation bulls were retained for analysis. For Methods 3 and 4, the numbers of SNPs deleted and retained were varied to obtain a maximum correlation between the GEBV and EBV of the validation bulls.

Correlations of the GEBV of the validation bulls with their current EBV were compared to the correlations of their PA with their EBV. Since the PA has a major effect on EBV of low heritability traits, even with >50 daughters, correlations of GEBV and PA with current MDYD were also computed. In addition, to estimate bias of PA and GEBV, relative to the current EBV, regressions of PA and GEBV on current EBV were computed, and means and standard deviations of PA, GEBV and current EBV were compared.

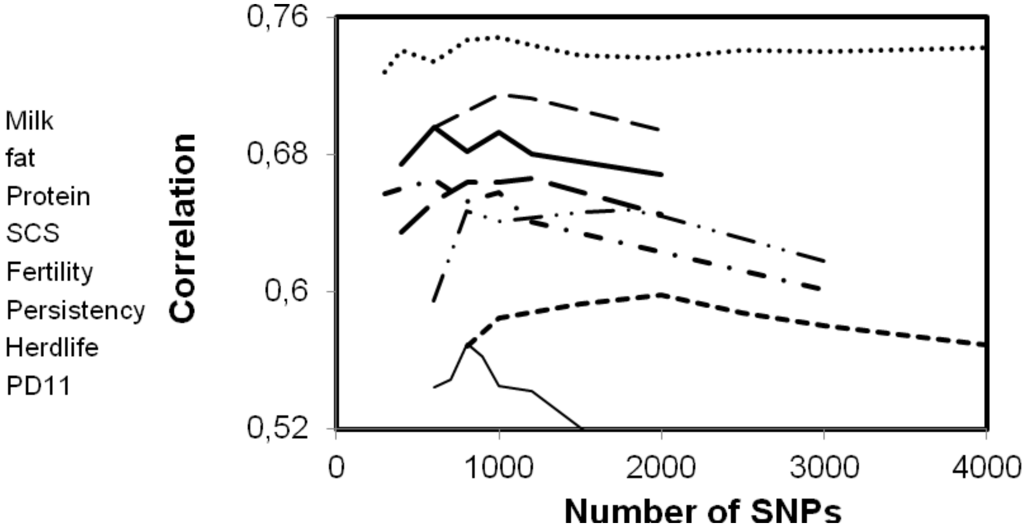
Results and discussion

Correlations of GEBV and PA with current EBV and MDYD from analysis of all SNPs and equally spaced SNPs are given in Table 2. With all valid SNPs, correlations of GEBV with current EBV and MDYD were slightly higher than parent averages for fertility, but lower for protein. With <2000 approximately evenly spaced SNPs, correlations of GEBV with current EBV were lower than PA for both traits.

Table 2. Correlations of GEBV and parent averages with current EBV and MDYD from analysis of all SNPs and equally spaced SNPs

		Correlations with current values			
Trait	No.	EBV		MDYD	
	SNPs	PA	GEBV	PA	GEBV
Protein	39,816	0.39	0.36	0.41	0.36
	1991		0.36		0.36
Fertility	39,816	0.66	0.67	0.36	0.41
	1991		0.62		0.37

Figure 2. Correlations between Method 1 GEBV and current EBV as a function of the numbers of SNPs included in the analysis.



The correlations between the Method 1 GEBV and current EBV as a function of the number of SNPs included in the analysis are plotted in Figure 2. There was a clear optimum for all the traits, except for fertility. The optimum number of markers was between 600 and 6000 for all of the traits analyzed.

Table 3. Correlations of Method 1 GEBV and parent averages with current EBV and MDYD with optimum number of SNPs.

		Correlations			
Traits	Optimum	EBV with:		MDYD with:	
	No. SNPs	PA	GEBV	PA	GEBV
Milk	600	0.55	0.69	0.56	0.67
Fat	1200	0.44	0.66	0.34	0.61
Protein	2000	0.39	0.59	0.41	0.57
SCS	600	0.53	0.66	0.42	0.60
Fertility	6000	0.65	0.73	0.36	0.49
Persistency	1000	0.60	0.71	0.45	0.62
Herdlife	800	0.37	0.54	0.17	0.35
PD11	1800	0.37	0.64	0.28	0.58

Correlations of Method 1 GEBV and PA with current EBV and MDYD with optimum number of SNPs are given in Table 3. The correlations of GEBV with current EBV and MDYD were higher than the correlations of PA with EBV and MDYD for all traits. The mean difference in the correlations between GEBV and PA was 0.18. Differences in correlations between PA and GEBV with current EBV and MDYD were similar for all traits. The greatest differences in correlations were obtained for PD11 for both EBV and MDYD, close to 0.3. Correlations between GEBV and EBV were higher for fertility and persistency, which have low heritability, due to the greater contribution of PA to EBV; while correlations of GEBV and PA with MDYD were lower. All correlations were lower for herd-life, which has only one record per cow. For protein the correlations of EBV with GEBV and PA were 0.59 and 0.39.

In Method 1, the SNPs were selected based on their November, 2011, EBV. In Method 2 SNPs were selected by the same procedure, but the dependent variables were the 2008 EBV. In this case GEBV were not more accurate than PA for any of the traits analyzed (data not shown). In the analysis of the validation set, the effect of PA was highly correlated with the sum of the SNP effects. Thus both the GEBV and PA detected the same QTL. However, as noted previously (Glick et al., 2012) the QTL segregating in the validation bulls are not the same as those segregating in the training population. In order to improve GEBV, it is necessary to include markers linked to QTL that are not segregating in the training set.

Correlations of Method 3 GEBV and PA with current EBV and MDYD with optimum number of SNPs are given in Table 4. The optimum number of markers deleted was 5000 for all traits, except protein. The optimum number of markers included ranged from 1000 to 1700. In nearly all cases the correlations of the GEBV with EBV and MDYD were higher with Method 3 than Method 1. The mean difference in the correlations between GEBV and parent averages was 0.26. Method 3 also used information not available in June, 2008.

The mean difference in the SNP allelic frequencies was 0.05, and the maximum difference was 0.3. Five percent of the SNPs (1786) had differences > 0.12. Correlations of Method 4 GEBV and PA with current EBV and MDYD with optimum number of SNPs are presented in Table 5. The optimum number of markers deleted ranged from 200 for fat to 8000 for persistency. The optimum number of markers included ranged from 800 for SCS to 9000 for persistency. For all traits, except for milk, the correlations of the GEBV with EBV and MDYD were higher than the correlations of parent averages with EBV and MDYD. The mean difference in the correlations between GEBV and parent averages was 0.042. This method, unlike Methods 1 and 3, only used information available in June, 2008.

Table 4. Correlations of Method 3 GEBV and parent averages with current EBV and MDYD with optimum number of SNPs.

			Correlations			
Traits	Optimum No. SNPs		EBV with:		MDYD with:	
	deleted	included	PA	GEBV	PA	GEBV
Milk	5000	1200	0.55	0.77	0.56	0.76
Fat	5000	1700	0.44	0.72	0.34	0.65
Protein	4000	1000	0.39	0.75	0.41	0.74
SCS	5000	1200	0.53	0.77	0.42	0.68
Fertility	5000	1200	0.65	0.79	0.36	0.53
Persistency	5000	1600	0.60	0.80	0.45	0.68
Herdlife	5000	1400	0.37	0.57	0.17	0.33
PD11	5000	1600	0.37	0.76	0.28	0.69

Table 5. Correlations of Method 4 GEBV and parent averages with current EBV and MDYD with optimum number of SNPs.

			Correlations			
Traits	Optimum No. SNPs		EBV with:		MDYD with:	
	deleted	included	PA	GEBV	PA	GEBV
Milk	500	2000	0.55	0.48	0.56	0.48
Fat	200	1000	0.44	0.52	0.34	0.43
Protein	1200	1500	0.39	0.47	0.41	0.48
SCS	1000	800	0.53	0.56	0.42	0.48
Fertility	1500	2000	0.65	0.66	0.36	0.39
Persistency	8000	9000	0.60	0.65	0.45	0.54
Herdlife	500	2000	0.37	0.43	0.17	0.21
PD11	800	1500	0.37	0.41	0.28	0.38

Table 6. Regressions and coefficients of determination of parent averages and Method 4 GEBV on EBV.

Traits	Regression on EBV		Coefficient of determination	
	PA	GEBV	PA	GEBV
Milk	1.00	0.79	0.30	0.23
Fat	0.85	1.13	0.20	0.27
Protein	0.87	1.06	0.16	0.22
SCS	0.85	0.76	0.28	0.31
Fertility	1.07	0.93	0.43	0.44
Persistency	0.96	1.12	0.36	0.42
Herdlife	0.75	0.86	0.14	0.19
PD11	0.89	1.11	0.13	0.17

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. Since true genetic values are unknown, GEBV and PA were compared to current EBV. Regressions and coefficients of determination of PA and Method 4 GEBV on EBV are presented in Table 6, and means and standard deviations of PA, Method 4 GEBV and current EBV are given in Table 7. For both PA and GEBV regressions were close to unity for all traits. With respect to means, GEBV were less biased than PA for milk production traits and PD11. Coefficients of determination for GEBV were higher than for PA for all traits except for milk production.

Table 7. Means and standard deviations of parent averages, Method 4 GEBV and current EBV.

Traits	Means			Standard deviations		
	PA	GEBV	EBV	PA	GEBV	EBV
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
Herdlife	55	60	51	41.3	42.2	83.5
PD11	466	377	370	127	124	335

Conclusions

GEBV derived from selected sets of markers can outperform GEBV derived from analysis of all markers. GEBV derived from selected sets of markers can outperform parent averages, even if the training population includes <1000 bulls. Using the optimum strategy correlations of GEBV with current EBV were 0.2 higher than correlations of PA with current EBV. Even if 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 correlations between GEBV and current EBV of 0.042 higher than correlations of PA with current EBV. Furthermore, GEBV were less biased than parent averages. It is likely that other methods of selection could improve upon these results.

Acknowledgements

This research was supported by grants from the Israel Dairy Board, the Chief Scientist of the Israeli Ministry of Agriculture & 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. We thank I. Misztal for use of the MTC REML program.

List of References

Glick, G., A. Shirak, S. Uliel, Y. Zeron, E. Ezra, E. Seroussi, M. Ron & J. I. Weller. 2012. Signatures of contemporary selection in the Israeli Holstein dairy cattle. Anim. Genet. (In press).

Habier, D., R. L. Fernando, & J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182: 343–353.

Moser, G., M. S. Khatkar, B. J. Hayes & H. W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42: 37.

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar & H. W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. **41**: 56.

Settar, P. & J. I. Weller. 1999. Genetic analysis of cow survival in the Israeli dairy cattle population. J. Dairy Sci. 82: 2170-2107.

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91: 4414-4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor & F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16–24.

Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, & D. B. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93: 5942–5949.

Weigel, K. A., G. de los Campos, O. Gonzalez-Recio, H. Naya, X. L. Wu, N. Long, G. J. Rosa, & D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires usingselected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92: 5248–5257.

Weller, J. I. & E. Ezra. 1997. Genetic analysis of somatic cell concentration and female fertility of Israeli Holsteins by the individual animal model. J. Dairy Sci. 80: 586-593.

Weller, J. I. & Ezra E. 2004. Genetic analysis of the Israeli Holstein dairy cattle population for production and nonproduction traits with a multitrait animal model. J. Dairy Sci. 87: 1519-1527.

Weller, J. I., E. Ezra & G. Leitner . 2006. Genetic analysis of persistency in the Israeli Holstein population by the multitrait animal model. J. Dairy Sci. 89: 2738-2746.

Zhang, Z., X. Ding, J. Liu, Q. Zhang, & D.-J. de Koning. 2011. Accuracy of genomic prediction using low-density marker panels. J. Dairy Sci. 94**:** 3642-3650.