Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Improving predictive ability of selected subsets of single nucleotide polymorphisms in a moderately sized dairy cattle population

J. I. Weller[1], E. Ezra[2], E. Seroussi[1], M. Shemesh[2], and M. Ron[1]
[1]ARO, the Volcani Center, [2]Israel Cattle Breeders Association

ABSTRACT: Three methods were tested to select subsets of markers from the Illumina BovineSNP50 BeadChip to compute genomic evaluations for moderately sized dairy cattle populations with ~1000 genotyped bulls. SNPs were selected based on: (1) their effects on the bulls' genetic evaluations for protein production in 2009 through 2013, as derived by the "EMMAX" algorithm including only bulls with daughter records; (2) their effects in the 2009 evaluations, also including bulls without daughter records; and (3) the regression of the SNPs allelic frequency on the bulls' birth dates. Milk, fat, and protein yield, somatic cell score, fertility, persistency, herd-life and the Israeli breeding index were analyzed by methods 2 and 3. Once SNPs were selected, only information available in 2009 was used to compute genomic evaluations for the validation bulls, who did not have daughter records in 2009. The optimum number of SNPs, as determined by correlations between genomic and 2013 evaluations, ranged from 800 for somatic cell score to 4000 for fertility. Correlations of up to 0.8 between genomic and 2013 evaluations for validation bulls were obtained by method 1 if markers were selected based on the 2013 evaluations, but in this case the selection criteria for SNPs is based on information not available in real time. Correlations of genomic with current evaluations greater than those obtained for parent averages were obtained by method 2 for most traits analyzed, and means were generally less biased than parent averages, thus method 2 is optimal for populations of this size.
Keywords: genomic evaluations; dairy cattle; genetics; single nucleotide polymorphisms

Introduction

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 (DYD) or deregressed estimated breeding values (EBV), and the independent variables are the genotypes of all valid SNPs. In nearly all cases, genomic EBV (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 GEBV of the validation bulls are then compared to their current EBV, DYD or deregressed EBV. With training populations consisting of thousands of bulls, reliabilities of > 0.7 can be obtained for young bulls with only pedigree and genotypic data (e. g., Wiggans et al. (2011)). However, if the training population consists of < 1000 bulls, reliabilities for GEBV are no greater than reliabilities based on parent average EBV (PA).

Correlations of GEBV of the bulls in the training set with current EBV are nearly always much higher than correlations of GEBV for the validation bulls. This is because linkage relationships and the frequencies of segregating quantitative trait loci change over time (Moser et al. (2009)). Glick et al. (2012) found that out of the 15,485 haplotypes with 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.

Considering the huge numbers of SNPs included on moderate and high density SNP chips, various studies have proposed computation of GEBV based on subsets of SNPs. Four basic strategies have been proposed to select SNPs: equally spaced SNPs throughout the genome; selection of SNPs with the greatest effects on the trait analyzed; selection of markers based on principal component analysis and selection of SNPs based on the difference in allelic frequencies of young and old bulls (reviewed by Weller et al. (2014)).

Weller et al. (2014) selected SNPs based on the effects of each marker on the bulls' genetic evaluations in 2012 and 2008, respectively, as derived by analysis of all valid SNPs by the "EMMAX" algorithm (Kang et al. (2010)). The difference between the correlation of GEBV and 2012 EBV and the correlation of PA and 2012 EBV was greater than 0.25 for all traits if SNPs were selected based on the 2012 evaluations, but not if SNPs were selected based on 2008 evaluations. Thus although it is possible to select subsets of markers that can be used to compute GEBV with reliabilities much higher than those computed based on all markers, the selection algorithm requires information not available in "real time."

Weller et al. (2014) also selected SNPs based on the differences in allelic frequency between the bulls in the training and validation sets. GEBV computed by SNPs selected by this criterion generally outperformed PA, but only marginally. Furthermore, GEBV were generally less biased than PA. In the current study two additional methods were tested for selecting SNPs based either on their

effects on the economic traits, or changes in allelic frequencies.

Materials and Methods

Animals genotyped and validation of SNPs. A total of 1132 bulls with reliabilities > 0.5 for milk production traits in May, 2013 were genotyped; 912 bulls for the 54,001 SNP BeadChip, and the remainder for the 54,609 SNP BovineSNP50 v2 BeadChip. Birth years ranged from 1975 through 2009.

SNPs were deleted from analysis if: they did not appear on the original BeadChip, the frequency of the less frequent allele less than 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 the first. After edits there were 38,556 valid SNPs.

The data set and traits analyzed. 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. EBV were computed for the complete data set (EBV_{13}), including all valid records from the Israeli Holstein population from January, 1985, through May, 2013, and the truncated data set including only records generated prior to June, 2009. 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 1-5 were included in the analyses. Herdlife was analyzed by a single trait model as described by Settar and Weller (1999).

The bulls with genotypes and EBV_{13} with reliabilities > 0.5 were divided into a "training set," of 884 bulls with reliabilities > 0.5 for milk production traits for EBV_{09} and the "validation set," of 163 bulls without daughter records in 2009 and reliabilities > 0.5 for production traits for EBV_{13}. The number of bulls in the two sets differed slightly for the non-production traits. 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 one year, and obtain their first EBV based on daughter records at approximately 5 years.

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 the truncated data set ($MDYD_{09}$), and for the validation bulls using all records generated up to May 2013 ($MDYD_{13}$), as described by Weller et al. (2014).

Selection of SNPs and evaluation of GEBV. Three methods were applied to select subsets of SNPs for analysis. In the first method, applied only to protein, SNPs were selected for each trait based on the fixed additive effect of each marker on the bulls' EBV for protein production in 2009 through 2013, as derived by analysis of all valid SNPs by the "EMMAX" algorithm. Five different subsets were selected, using the EBV computed in June of each year. Only bulls with EBV based on daughter records were used to determine SNP effects. In the second method, which was applied to all eight traits, SNPs were selected based on their effects in the June, 2009 evaluation, but for validation bulls without EBV based on daughter records, EBV were computed as PA based on the 2009 EBV of the parents. In the third method, also applied to all eight traits, SNPs were selected based on the regression of the SNPs allelic frequency on the bulls' birth dates, provided that the frequency of the less frequent allele was > 0.1 for bulls born after 1984 with EBV in 2009. The SNPs with the greatest absolute values for the regression were retained.

For all three methods in the preliminary analysis the 1000 SNPs with the greatest effects, were selected for inclusion. The number of SNPs included was then increased by increments of 500 up to 6000, or until a decrease of greater than 2% in the correlation of the GEBV of the validation bulls with their EBV_{13} was obtained. In each additional run, the 500 SNPs with the next greatest effects were added to the previous sample of SNPs. 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.

Computation and evaluation of GEBV. The method of VanRaden (2008) was used to compute genomic effects on the MDYD from the training set for each trait. All bulls with genotypes and $MDYD_{09}$ were included in the analysis. Direct genomic evaluations (DGE) for the validation bulls were then computed as: $\mathbf{Z}\hat{\mathbf{a}}$, where $\mathbf{Z}$ is the incidence matrix that relates $MDYD_{09}$ with the genomic effects vector $\mathbf{a}$ and $\hat{\mathbf{a}}$ is the vector of solutions for $\mathbf{a}$. Similar to VanRaden et al. (2009) final GEBV were computed from an index including the DGE and PA. The regression coefficients for the index were derived from the training data set. PA were computed as the means of the parent EBV_{09} derived from the standard multi-trait analysis of the truncated data set. PA and GEBV derived by the three methods were compared to EBV_{13} and $MDYD_{13}$ based on correlations, regressions of EBV_{13} on GEBV and means and standard deviations (SD).

Results and Discussion

Correlations of PA with EBV_{13} and $MDYD_{13}$, PA means, SD and regressions of EBV_{13} on PA are given in Table 1 for all 8 traits. Correlations were generally higher for EBV_{13}, than for $MDYD_{13}$, because parent evaluations contribute to the EBV_{13} through the relationship matrix. The difference was more pronounced for the low heritability traits. Evaluations can be considered unbiased if regressions are close to unity and means of PA are close to the

EBV_{13} means. Regressions were close to unity for all traits, except for herdlife and PD11. Means of PA were higher than EBV_{13} means for all traits, except for SCS and fertility. The greatest difference in SD units was obtained for milk, nearly 0.5 SD.

Table 1. Correlations of parent average EBV (PA) based on 2009 data, with 2013 EBV (EBV_{13}) and modified daughter-yield deviations (MDYD) for each trait; PA and EBV means, PA standard deviations (SD) and regressions of EBV_{13} on PA (Reg).

	Correlations		Means			
Trait	EBV_{13}	MDYD	PA	EBV_{13}	SD	Reg
Milk	0.54	0.55	348.9	256.7	189	1.00
Fat	0.50	0.41	16.64	14.84	7.90	0.91
Protein	0.47	0.47	15.16	14.28	4.20	0.99
SCS[1]	0.61	0.51	-0.08	-0.07	0.12	0.9
Fertility	0.60	0.34	0.13	0.34	1.42	1.04
Persistency	0.62	0.54	0.64	0.2	1.46	0.93
Herdlife	0.50	0.05	70	64.5	52.2	0.76
PD11	0.43	0.40	532.0	465.2	165	0.82

[1] Somatic cell score

Optimum numbers of SNPs, correlations, means, SD and regressions of protein EBV_{13} on GEBV derived by method 1 are presented in Table 2. With effects selected based on 2009 EBV, correlations of GEBV with EBV_{13} and $MDYD_{13}$ were lower than the corresponding correlations with PA. However, with selection based on 2010 EBV correlations with EBV_{13} and $MDYD_{13}$ were higher than the corresponding PA correlations. With selection based on 2013 EBV, correlations approached 0.8, as found previously by Weller et al. (2014). In all cases, data collected after 2009 was used only to select SNPs included in the analysis. The calculation of the GEBV used only data available in 2009. These results indicate that reliabilities of GEBV derived by method 1 should be higher for older calves. This was tested by computing correlations by birth year, but the results were not consistent, apparently due to the relatively low number of bulls genotyped each year. Unlike PA, the method 1 estimates had regressions < unity for the early years, and regressions > unity for the later years. There were no clear trends for means or SD. All SD were higher than for PA.

Table 2. Optimum numbers of SNPs, correlations of GEBV derived by subsets of SNPs selected by EMMAX effects with EBV computed in 2009 through 2013, with 2013 EBV (EBV_{13}) and modified daughter-yield-deviations (MDYD); means and standard deviations (SD) of GEBV and regressions of 2013 protein EBV on GEBV (Reg).

	Number	Correlations				
Year	of SNPS	EBV_{13}	MDYD	Mean	SD	Reg
2009	500	0.40	0.40	9.23	5.72	0.61
2010	1000	0.53	0.54	10.00	5.91	0.79
2011	500	0.68	0.68	11.08	6.06	0.98
2012	4000	0.70	0.71	10.39	5.62	1.10
2013	4000	0.78	0.77	10.64	5.70	1.20

Correlations of PA and GEBV derived by method 2 with EBV_{13} and $MDYD_{13}$ for the optimum number of SNPs for each trait are given in Table 3. The optimum number of SNPs ranged from 800 for SCS to 4000 for fertility. For all traits, except for persistency and PD11, the correlations of $MDYD_{13}$ with GEBV were higher than the correlations with PA. This was also the case for correlations of GEBV with EBV_{13} for the milk production traits. SD were greater for all traits by method 2, but all regressions were < 1, except for fat. Means were generally less biased than PA.

Table 3. Correlations of method 2 GEBV with 2013 EBV (EBV_{13}) and modified daughter-yield deviations (MDYD) for the optimum number of SNPs for each trait; means and standard deviations (SD) of method 2 EBV and regressions of EBV_{13} on method 2 EBV (Reg).

Trait	Number	Correlations				
	of SNPs	EBV_{13}	MDYD	Means	SD	Reg
Milk	1500	0.57	0.58	305.1	206	0.96
Fat	2500	0.57	0.51	14.28	7.55	1.07
Protein	1500	0.48	0.50	11.7	4.77	0.89
SCS[1]	800	0.61	0.57	-0.11	0.153	0.75
Fertility	4000	0.61	0.38	0.35	1.68	0.9
Persistency	3000	0.59	0.50	0.715	1.68	0.78
Herdlife	1000	0.47	0.09	81.0	61.3	0.67
PD11	2000	0.38	0.39	434.2	167.7	0.7

[1] Somatic cell score

Correlations of method 3 GEBV with EBV_{13} and $MDYD_{13}$ were at best only marginally better than PA, and are therefore not presented.

Conclusions

Three methods were tested to select subsets of markers to compute GEBV for moderately sized dairy cattle populations with ~1000 genotyped bulls. The optimum number of SNPs ranged from 800 for SCS to 4000 for fertility. Correlations of up to 0.8 between GEBV and current EBV for validation bulls can be obtained using selected subsets of markers, but only if the selection criteria is based on information not available in real time. Correlations of GEBV with current EBV greater than those obtained between PA and current EBV can be obtained for most traits analyzed with subsets of markers selected based on their effects as derived by the EMMAX algorithm. Means were generally less biased than PA. Thus this method is optimal for populations of this size.

Literature Cited

Glick, G. Shirak, A., Uliel, S. et al. (2012). Anim. Genet. 43 (Suppl. 1):45-55.

Kang, H. M., Sul, J. J., Service, S. K. et al. (2010). Nature Genet. 42:348-354.

Moser, G., Khatkar, M. S., Hayes, B. J. et al. (2010). Genet. Sel. Evol. 42:37.

Settar, P., and Weller, J. I. (1999). J. Dairy Sci. 82:2170-2177.

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

VanRaden, P. M, Van Tassell, C. P., Wiggans, G. R. et al. (2009). J. Dairy Sci. 92:16-24.

Weller, J. I., and Ezra, E. (2004). J. Dairy Sci. 87:1519-1527.

Weller, J. I., Ezra, E., and Leitner G. (2006) J. Dairy Sci. 89:2738-2746.

Weller, J. I., Glick, G., Shirak, A., et al. (2014). Animal 8:208-216.

Wiggans, G. R., VanRaden, P. M., and Cooper, T. A. (2011). J. Dairy Sci. 94:3202-3211.