Animal (2012), **6:2**, pp 193–202 © The Animal Consortium 2011
doi:10.1017/S1751731111001704

animal

A simple method for genomic selection of moderately sized dairy cattle populations

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

[1]*Department of Ruminant Science, Institute of Animal Sciences, A.R.O., The Volcani Center, Bet Dagan 50250, Israel;* [2]*Department of Ruminant Science, The Robert H. Smith Faculty of Agriculture, The Hebrew University of Jerusalem, Rehovot, Israel;* [3]*Department of Ruminant Science, Sion – Israeli Company for Artificial Insemination & Breeding Ltd, M. P. Shikmim 79800, Israel;* [4]*Department of Ruminant Science, Israel Cattle Breeders Association, Caesarea Industrial Park 38900, Israel*

(Received 16 November 2010; Accepted 18 July 2011; First published online 26 September 2011)

An efficient algorithm for genomic selection of moderately sized populations based on single nucleotide polymorphism chip technology is described. A total of 995 Israeli Holstein bulls with genetic evaluations based on daughter records were genotyped for either the BovineSNP50 BeadChip or the BovineSNP50 v2 BeadChip. Milk, fat, protein, somatic cell score, female fertility, milk production persistency and herd-life were analyzed. The 400 markers with the greatest effects on each trait were first selected based on individual analysis of each marker with the genetic evaluations of the bulls as the dependent variable. The effects of all 400 markers were estimated jointly using a 'cow model,' estimated from the data truncated to exclude lactations with freshening dates after September 2006. Genotype probabilities for each locus were computed for all animals with missing genotypes. In Method I, genetic evaluations were computed by analysis of the truncated data set with the sum of the marker effects subtracted from each record. Genomic estimated breeding values for the young bulls with genotypes, but without daughter records, were then computed as their parent averages combined with the sum of each animal's marker effects. Method II genomic breeding values were computed based on regressions of estimated breeding values of bulls with daughter record on their parent averages, sum of marker effects and birth year. Method II correlations of the current breeding values of young bulls without daughter records in the truncated data set were higher than the correlations of the current breeding values with the parent averages for fat and protein production, persistency and herd-life. Bias of evaluations, estimated as a difference between the mean of current breeding values of the young bulls and their genomic evaluations, was reduced for milk production traits, persistency and herd-life. Bias for milk production traits was slightly negative, as opposed to the positive bias of parent averages. Correlations of Method II with the means of daughter records adjusted for fixed effects were higher than parent averages for fat, protein, fertility, persistency and herd-life. Reducing the number of markers included in the analysis from 400 to 300 did not reduce correlations of genomic breeding values for protein with current breeding values, but did slightly reduce correlations with means of daughter records. Method II has the advantages as compared with the method of VanRaden in that genotypes of cows can be readily incorporated into the Method II analysis, and it is more effective for moderately sized populations.

Keywords: genomic selection, quantitative trait loci, single nucleotide polymorphisms, dairy cattle breeding

Implications

Two algorithms for genomic selection of moderately sized populations are described, and applied to the analysis of milk, fat and protein production; somatic cell score (SCS); female fertility; persistency; and herd-life of Israeli Holsteins. Method II genomic evaluations were less biased and more accurate than parent averages for fat, protein, female fertility, persistency and herd-life, but not for milk production and SCS. The bias for milk production traits was slightly negative, as opposed to the positive bias of parent averages. The proposed method can be readily applied to the traits analyzed by animal models, and is effective for moderately sized populations, and can incorporate genotypes from both bulls and cows.

Introduction

Genome scans based on thousands of single nucleotide polymorphisms (SNPs) covering the entire genome have been completed or are in progress for several dairy cattle populations (Cromie *et al.*, 2010). Implementation of genomic selection

† E-mail: weller@agri.huji.ac.il

requires combining marker information with pedigree and phenotypic data in order to obtain genomic estimated breeding values (GEBVs). Goddard and Hayes (2007) considered three alternatives, the second of which was to infer marker genotypes for all animals, and to use these to calculate GEBV.

Most studies that have compared GEBVs with estimated breeding values (EBVs) without marker data have done so by dividing the population of bulls into two groups (e.g. VanRaden *et al.*, 2009a). The older bulls, called the 'predictor group,' are used to derive estimates of the marker effects. These estimates are then used to derive GEBVs for the second group of young bulls, based only on marker and pedigree data. The GEBVs of the young bulls are then compared with their current EBVs based on daughter records.

VanRaden (2008) proposed analysis of daughter yield deviations (DYD; VanRaden and Wiggans, 1991) as the dependent variable, with all SNP markers included as random effects. This model requires weighting the residuals as a function of the DYD reliabilities. Genotypes for 38 416 informative markers and the August 2003 genetic evaluations for 3576 Holstein bulls born before 1999, were used to predict the January 2008 daughter deviations for 1759 bulls born between 1999 and 2002. Predictions were computed using linear and nonlinear genomic models. For linear predictions, the traditional additive genetic relationship matrix was replaced by a genomic relationship matrix, which is equivalent to assigning equal genetic variance to all markers. For nonlinear predictions, markers with smaller effects were regressed further toward zero; markers with larger effects were regressed less to account for a non-normal prior distribution of marker effects (VanRaden, 2008). The final genomic predictions combined three terms by the selection index:

1. direct genomic prediction,
2. parent averages computed from the set of genotyped ancestors using traditional relationships,
3. published parent averages or pedigree indexes, constructed as 0.5 (sire EBVs) + 0.25 (maternal grandsire EBVs) + 0.25 (birth year mean EBVs).

For each animal, a 3×3 matrix was set up with reliabilities for the three terms on the diagonals and functions of these reliabilities on the off-diagonals. Differences between the linear and nonlinear prediction models were minimal.

Most countries implementing genomic selection are using similar methodologies based on analysis of either DYD or 'deregressed EBV' (e.g. Ducrocq *et al.*, 2009; Loberg and Durr, 2009). Advantages of this system for genomic evaluation include no change to the regular evaluations and simple steps for predicting genomic values for young genotyped animals. Disadvantages of this system include estimation of weighting parameters, such as variance components (Guillaume *et al.*, 2008) or selection index coefficients (VanRaden *et al.*, 2009a), loss of information and biased evaluations (Aguilar *et al.*, 2010). If genomic selection is used, the expectation of Mendelian sampling in selected animals is not zero (Party and Ducrocq, 2009).

The method of VanRaden has been shown to work well only for very large data sets consisting of thousands of genotyped bulls (VanRaden *et al.*, 2009a and 2009b). A joint evaluation using all phenotypic, pedigree and SNP data should be the optimal strategy (Aguilar *et al.*, 2010). However, application is problematic because of the huge number of equations required, and the fact that only a very small fraction of the population, chiefly bulls, is genotyped.

Legarra *et al.* (2009) and Misztal *et al.* (2009) proposed analysis of SNPs as random effects, with conditioning of the genetic value of ungenotyped animals on the genetic value of genotyped animals via the pedigree information, and then to use the genomic relationship matrix for the latter. This results in a joint distribution of genotyped and ungenotyped genetic values, with a pedigree–genomic relationship matrix, **H**. This method was applied to 10 466 066 US Holsteins' records for final score (Aguilar *et al.*, 2010). The GEBVs were computed based on 6508 bulls. This approach includes a parameter, λ, that is related to the fraction of the additive variance explained by the genomic information. This parameter is the population–genetic equivalent of variance components in traditional mixed model analyses. With 'optimal' scaling, the regression of GEBVs on EBVs based on daughter records using the method of Aguilar *et al.* (2010) was 0.90, as opposed to 0.83 by the multi-step method of VanRaden (2008), although both methods had nearly equal coefficient of determination (R^2) values. Similar results were obtained on five additional conformation traits (Tsuruta *et al.*, 2010).

Israel and Weller (1998) proposed a model in which the marker effects were included in the complete animal model (AM) analysis as fixed effects. For animals that are not genotyped, probabilities of receiving either allele were included as regression constants. Although on simulated results the model was able to derive unbiased estimates of quantitative trait loci (QTL) effects, on real data the QTL effects were underestimated, relative to alternative estimation methods (Weller *et al.*, 2003). Baruch and Weller (2009) found that bias increased as a function of the number of generations included in the analysis, as the fraction of animals genotyped decreased, and as the QTL allelic frequencies became more extreme. They concluded that the main reason for this bias was due to confounding between the QTL effect and the polygenic effect, as estimated via the relationship matrix. They proposed a modified 'cow model' that does not account for relationships among animals. Using this model, they were able to derive unbiased estimates for QTL effects on simulated data, even though only a small fraction of the population was genotyped.

Although this model was able to derive unbiased estimates of QTL effects, it could not be used for ranking animals for selection, because it does not include the relationship matrix, and only cows with records were included in the analysis. They therefore proposed a 5-stage algorithm: the QTL substitution effects are estimated by the modified cow model; the phenotypic records are then adjusted by subtraction of the appropriate QTL effects for each animal. Next, the adjusted records were analyzed by a standard multiple

trait AM. Finally, the QTL effects of each animal were added to the AM evaluation, and these evaluations were used to rank animals for selection. On simulated data with one or two segregating QTL, this method was able to derive unbiased genetic evaluations and genetic progress was increased relative to a standard AM (Baruch and Weller, 2008 and 2009).

Generally, the basis for comparison between the young bull evaluations based either only on pedigree or pedigree and marker data and the EBVs of these bulls based on daughter records is the R^2. GEBVs by the method of VanRaden (2008) were more accurate than official parent averages for all 27 traits analyzed. R^2 of GEBVs for EBVs based on daughter records were 0.02 to 0.38 higher with nonlinear genomic predictions included as compared with parent averages alone. However, gains were much lower for smaller numbers of predictor bulls, and were nearly linear functions of the number of predictor bulls (VanRaden *et al*., 2009a). With only 1151 predictor bulls, the gain in the R^2 of daughter deviations was only 4% for net merit. Thus, this method is not appropriate for analysis of moderately sized populations.

A second important criterion that should be considered is the bias of GEBVs. That is, GEBVs are unbiased if the regression of true breeding values on GEBVs is not significantly different from unity, and if the *y*-intercept is not significantly different from zero (Mäntysaari *et al*., 2010). If the regression is less than unity, then the evaluations of the bulls with the highest GEBVs will be inflated relative to the true genetic values of these bulls. Aguilar *et al*. (2010) found that regressions of EBVs based on progeny tests on GEBVs, derived by the method of VanRaden (2008), were less than unity for the final score. Liu *et al*. (2009) developed a method for correcting for bias generated by the selection of animals for genotyping.

The objectives of this study were to apply the method of Baruch and Weller (2008 and 2009) to the actual genotype data of the Israeli Holstein population for the BovineSNP50 BeadChip (Matukumalli *et al*., 2009), using the 400 markers with the largest effects, in order to derive GEBVs for production and nonproduction economic traits, and to evaluate the GEBV based on R^2 and bias.

Material and methods

Genomic data

A total of 1143 Israeli Holstein bulls have been genotyped for the 54K Beadchip. Of these, 912 bulls were genotyped for the BovineSNP50 BeadChip, which contains 54 001 genetic markers, and 265 bulls were genotyped for BovineSNP50 v2 BeadChip, which contains 54 609 genetic markers. Of these, 916 bulls have EBVs for milk production traits based on daughter records. Twenty-five bulls genotyped for the original BovineSNP50 BeadChip were also genotyped for BovineSNP50 v2 BeadChip. SNPs were deleted from further analysis if there were valid genotypes for less than half the bulls analyzed, or if the frequency of the more common allele was >0.95. If the identity of genotypes for two consecutive SNPs on the same chromosome was >0.95, then the second SNP was deleted. This left 40 094 valid SNPs as compared with 38 416 in the US analysis. There were 670 bulls for which the father was also genotyped, and paternity was verified for these bulls. For a diallelic marker, if only the progeny and a single parent are genotyped, there is a conflict only if the progeny and putative parent are homozygous for different alleles. There were 24 bulls with $>4.5\%$ conflicts between the genotypes of the putative sire and son. Thus, assuming all DNA samples were correctly identified, the frequency of incorrectly recorded paternity was 3.6%, even though paternity was previously validated by microsatellites for nearly all of these bulls. Excluding these 24 bulls with incorrect paternity, all other discrepancies between sire and son genotypes were assumed to be due to genotype mistakes (Weller *et al*., 2010).

The economic traits analyzed

The traits analyzed were milk, fat and protein production; somatic cell score (SCS); female fertility; milk production persistency; and herd-life. Female fertility was defined as the inverse of the number of inseminations to conception. Herd-life was analyzed by a single-trait AM as described by Settar and Weller (1999). All the other traits were analyzed by the following multitrait AM, with each parity considered a separate trait (Weller and Ezra, 2004; Weller *et al*., 2006):

$$Y_{ijklm} = H_{ij} + \mathrm{PA}_{im} + G_{ik} + A_{il} + \mathrm{PE}_{il} + e_{ijklm} \qquad (1)$$

where Y_{ijklm} is the record for parity m of cow l from herd-year-season (HYS) j for trait i, H_{ij} the fixed effect of HYS j on trait i, PA_{im} the fixed effect of parity m for trait i, G_{ik} the effect of genetic group k on trait i, A_{il} the random additive genetic effect of cow l for trait i, PE_{il} the random permanent environmental effect of cow l for trait i and e_{ijklm} the random residual. Records were pre-adjusted for calving age and month and days open as described previously (Ezra *et al*., 1987). EBVs were computed for all cows and bulls in the entire Israeli Holstein population, based on all valid first through fifth parity cow records since 1985. The genetic base for all traits was the mean of cows born in 2005.

In the truncated data set, all records with freshening dates after September 2006, were deleted. The numbers of cows, bulls and HYSs included in the truncated and complete data sets for each trait are given in Table 1. Depending on the trait analyzed, there were 148 to 192 'young bulls' with genotypes and EBVs based on daughter records in the complete data set, but without daughter records in the truncated data set. The number of bulls with genotypes and EBVs in the complete data set, and the number of young bulls with genotypes, but without EBVs in the truncated data set, are given in Table 2. Foreign bulls and bulls of breeds other than the Holstein were excluded, because these bulls do not accurately represent the Israeli Holstein population. Mean daughter deviations (MDD) per sire for the young bulls were computed as the weighted mean of sire's daughter records, with the appropriate HYS and parity effects subtracted from the phenotypic record. Means of cow records were weighted

by $n/(n+\delta)$ where n is the number of records per cow and δ is the ratio of residual to cow effect variance. For milk production traits, $\delta = 1$. MDD differ from DYD in that MDD are not corrected for merit of mates.

Selection of markers for inclusion in genomic evaluation

Significant linkage disequlibrium between SNP genotypes of the bulls and QTL for all 40 094 valid SNPs for all traits analyzed was estimated by the following regression model:

$$EBV_{ij} = S + BY + BY^2 + e_{ij} \quad (2)$$

where EBV_{ij} is the breeding value of bull i for trait j, S the SNP genotype, BY the birth year and e_{ij} the random residual. Heterozygotes were scored as 1 and homozygotes as either 0 or 2.

For each trait, the 2000 SNPs with the lowest P-values for the effect of SNP genotype were reanalyzed individually by the MTC restricted maximum likelihood (REML) program, for the following model:

$$EBV_{ij} = S + A_{ij} + e_{ij} \quad (3)$$

where A_{ij} is the additive genetic effect of bull i for trait j, and the other terms are as described previously. The additive genetic and SNP effects were considered random variables in this analysis. The relationship matrix included all known parents and grandsirs of the bulls with genotypes. Although each individual has two parents and two grandsires, not all female ancestors were known, and nearly all the male ancestors were already included among the genotyped animals. Therefore, inclusion of these ancestors approximately doubled the number of animals included in the analysis. The allelic substitution effect was derived under the assumption of additivity from the following equation:

$$r_{vj} = \sigma_{sj}/[2p_s(1-p_s)]^{0.5} \quad (4)$$

where r_{vj} is the allelic substitution effect for trait j, σ_{sj} the square root of the SNP component of variance for trait j and p_s the frequency of the less frequent marker allele in the population of bulls.

Significance values for the SNP component of variance were determined by permutation analysis. One thousand repeat analyses of a typical data set with randomization of the SNP effects relative to the bulls' EBVs were run. The SNP component of variance was >2% of the total variance for <5% of the repeat samples. Thus, this criterion was used to determine significance of the marker effect.

Analysis methods and selection schemes

For each trait analyzed, the 400 markers with the greatest variance components for that trait were used for computation of GEBVs. Thus, a different set of markers were analyzed for each trait. Method I GEBVs for the young bulls were computed by the following algorithm:

1. For animals that were genotyped, we assumed that all genotypes were determined without error. Probabilities of genotypes for all other animals were computed based on the algorithm of Kerr and Kinghorn (1996). For animals with unknown parents, genotype probabilities were assumed to be equal to the mean probabilities in the entire sample of genotyped bulls.
2. The effects of these 400 markers were estimated jointly by the 'cow model' of Baruch and Weller (2008).
3. The sum of the 400 marker effects as estimated by the cow model were subtracted from the production records of the cows based on each cow's genotype probabilities, and the marker effects as estimated by the cow model.

Table 1 *Numbers of cows, bulls and HYS included in the truncated and complete data sets for each trait*

	Truncated data set			Complete data set		
Trait	Cows	Bulls	HYS	Cows	Bulls	HYS
Production[1]	661 548	1836	31 401	780 520	2284	37 029
SCS	565 400	1711	28 085	682 672	2156	33 620
Fertility	631 287	1854	34 896	739 307	2274	40 592
Persistency[2]	614 133	1774	63 562	733 105	2213	74 586
Herd-life	731 322	1936	37 669	827 381	2300	42 077

HYS = herd-year-seasons; SCS = somatic cell score.
[1]Numbers of animals and HYS were the same for milk, fat and protein production.
[2]Numbers of HYS were greater for persistency, because first and later parities were assigned to separate HYS.

Table 2 *Regressions of the EBV on the sum of their marker effect values*

	Number of bulls					R^2	
Trait	All	Young	h^2	Mean reliability	Regression±s.d.	All bulls	Young bulls
Milk (kg)	916	186	0.25	0.94	0.62 ± 0.03	0.35	0.10
Fat (kg)	916	186	0.30	0.94	0.78 ± 0.04	0.34	0.11
Protein (kg)	916	186	0.25	0.94	0.98 ± 0.04	0.45	0.07
SCS	916	192	0.15	0.90	0.20 ± 0.02	0.10	0.02
Fertility (%)	907	183	0.02	0.83	0.24 ± 0.02	0.19	0.02
Persistency (%)	969	191	0.20	0.91	0.62 ± 0.02	0.44	0.15
Herd-life	979	148	0.11	0.84	0.67 ± 0.03	0.39	0.15

EBV = estimated breeding values; SCS = somatic cell score.

4. Multitrait AM evaluations were then computed for the adjusted records of the truncated data set. These EBVs are now based only on the effects not accounted for by the markers.
5. The GEBVs of the young bulls were computed as the parent average of the EBVs, which were computed from the adjusted records plus the sum of marker effects for each young bull.

The 'cow model' used to estimate the marker effects was as follows:

$$Y_{ijklmn} = C_{in} + H_{ij} + \mathrm{PA}_{ik} + \sum_{m=1}^{M} q_{lm} r_{im} + e_{ijklmn} \quad (5)$$

where Y_{ijklmn} is the record of cow n in parity k for trait i, C_{in} the random effect of cow n for trait i, q_{lm} the inferred genotype probability l for marker m based on each cow's genotyped ancestors, r_{im} the effect of marker m on trait i; and the other terms are as defined for equation (1). q_{lm} were scored over the scale of 0 to 1, where 0 = homozygote for the 'negative' QTL allele and 1 = homozygote for the 'positive' QTL allele. Marker substitution effects were computed as $0.5r_{im}$. As only sires were genotyped, no female could have an inferred genotype of either 0 or 1. Covariances among the random cow effects are assumed to be zero. That is, the relationship matrix was not included. Although the cow effect includes the polygenic and the permanent environmental effects, it does not include the QTL effects. As the magnitude of these effects is not known, the cow model was first run under that assumption of QTL effects of zero. That is, the variance due to the cow effect is equal to 0.5 for milk production traits. Once estimates of the marker effects were obtained, the sum of the variances due to the marker effects were deleted from the cow effect as follows:

$$\sigma_c^2 = \sigma_i^2 - \sum [p_m(1-p_m)r_m^2] \quad (6)$$

where σ_c^2 is the variance of the cow effect after subtraction of QTL effects, σ_i^2 the variance of the cow effect including QTL effects, p_m the estimated frequency of the less frequent allele for QTL i and r_m the estimated substitution effect for QTL m. The cow model was run three times with estimates of σ_c^2 updated at each iteration.

In Method II, regression coefficients were derived from the predictor bull population as follows:

$$\mathrm{EBV}_{ijk} = \mathrm{int}_i + a \times \mathrm{MEBV}_{ij} + b \times \mathrm{MS}_{ij} + c \times B_k + e_{ijk} \quad (7)$$

where EBV_{ijk} is the EBV of bull j for trait i in the truncated data set, int_i the y-intercept for trait i, MEBV_{ij} the mean of bull's parents EBV in the truncated data set for trait i, MS_{ij} the sum of marker scores for this bull for trait i, B_k the bull's birth year k, e_{ijk} the random residual, and a, b and c are regression coefficients. GEBVs for the young bulls were then derived from equation (7) using the y-intercept and regression constants obtained from the predictor bulls. As a was generally <1 and b was always <1, this method can be considered more 'conservative' than Method I, in that standard deviations among the evaluations were smaller. The GEBVs were also computed by the method of equation (1) in VanRaden (2008) for protein production (Method III). The dependent variable was the sires' MDD, as described previously. Only 730 bulls with reliabilities >50 in the truncated data set were included. Residual variances were derived from the sires' reliabilities, as described by VanRaden (2008). All 40 094 valid markers were included. The GEBVs were computed as Zû. The Zû for the young bulls were then computed based on the marker effects derived from the truncated data set.

The GEBVs by all three methods were compared with the current genetic evaluations for bias, computed as the difference in the means and regression of current EBVs on GEBVs. Correlations were also computed between the current evaluations and the three GEBV methods and parent averages computed from the truncated data set. Finally, correlations were computed between MDD, GEBV, parent averages and current EBV.

Results

The F-values, the nominal P-values and the false discovery rate (FDR; Weller *et al.*, 1998) values for the 50 SNPs with the lowest P-values by the regression model in equation (2) are given in Table 3 for the traits analyzed. All FDR values were $\leqslant 0.1$, and thus nearly all of these effects can be considered

Table 3 P-*values and FDR levels for the 50 SNP effects with the lowest* P-*values for the traits analyzed*

Trait	*F*-value	Nominal *P*-values	Expected number of significant effects	FDR	SNPs with significant REML variance component
Milk	20.8	5.8×10^{-6}	0.23	0.005	1320
Fat	22.2	2.8×10^{-6}	0.11	0.002	1327
Protein	24.5	8.9×10^{-7}	0.03	7.0×10^{-4}	1129
SCS	23.8	1.2×10^{-6}	0.05	9.8×10^{-4}	717
Fertility	45.4	2.7×10^{-11}	1.1×10^{-6}	2.2×10^{-8}	1123
Persistency	23.4	3.2×10^{-4}	12.9	0.013	1277
Herd-life	31.1	3.1×10^{-8}	0.0013	2.7×10^{-5}	1872

FDR = false discovery rates; SNP = single nucleotide polymorphism; REML = restricted maximum likelihood; SCS = somatic cell score.

'real' effects, unless the SNP effect is due to confounding with the polygenic effect due to relationships, as demonstrated by Habier *et al*. (2007). Surprisingly, with the exception of persistency, there is an inverse relationship between the trait heritability and the FDR values; the higher the trait heritability, the lower the number of genes with detectable effects. From 717 (for SCS) to 1872 (for herd-life) of the 2000 effects with the lowest *P*-values by the linear model also had significant marker effects by the REML analyses. That is, the marker explained >2% of the variance. Thus, overall, more than half of the SNPs with the greatest effects by the regression model were also significant by the variance component model. The value for SCS was much lower than for the other traits. There was higher correspondence between the linear and REML models for the production traits, which had higher heritability. The correlations between the regression model *F*-values, derived from equation (2) and the SNP variance component derived from the 2000 REML analysis (equation (3)) were generally in the range of 0.2, but only 0.1 for protein (data not shown).

Comparison of SNP substitution effects for protein from the REML analysis (r_{vj}, equation (4)) and the cow model (r_{im}, equation (5)) are given in Figures 1 and 2. In Figure 1, the cow model effects are plotted as a function of the REML effects. The regression line is also plotted. As expected, the cow model effects are smaller because all the effects were included in the same analysis, whereas in the REML analyses each SNP was analyzed separately. The sum of the marker variances from the individual REML analyses explains more than 100% of the total sire variance. This is not surprising, as several markers may be linked to the same QTL. Furthermore, REML effects are biased upward. In addition, the SNP effects were considered random in the REML analysis and fixed in the cow model. Ramifications of whether these effects are considered fixed or random were considered in detail by Gianola *et al*. (2009). The correlation between the SNP effects for protein production from REML analysis was only 0.36. The SNP with the greatest effect in the cow model was ranked 15th by the REML analysis, and the SNP with the third largest effect by the AM was ranked 121th by the REML analysis.

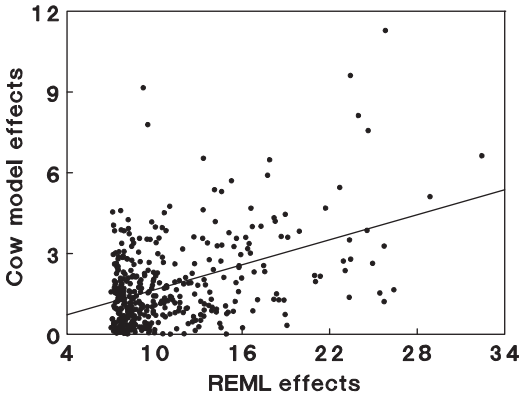

Figure 1 The cow model substitution effects as a function of the restricted maximum likelihood substitution effects for kg protein. The regression line is also plotted.

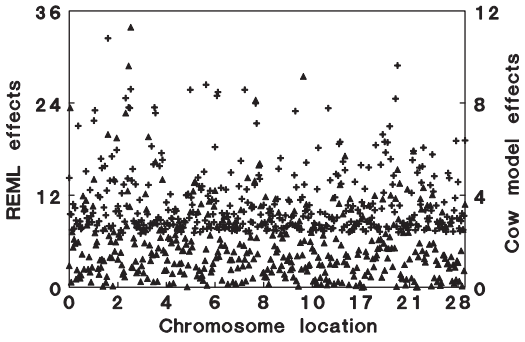

Figure 2 Comparison of single nucleotide polymorphism substitution effects for protein from the restricted maximum likelihood (REML) analysis and the cow model ordered by chromosomal location. +, REML effects; ▲, cow model effects.

In Figure 2, SNP effects for protein from the REML analysis and the cow model are plotted by chromosomal location. Although one of the largest REML effects was found on *Bos taurus* chromosome (BTA) 6 near the location of *ABCG2* (Cohen-Zinder *et al*., 2005), this was not the case for the cow model effects. The largest REML effects were found on BTA1 and 20, although no major QTL affecting protein has been verified in these regions. The largest cow model effects were obtained on BTA2 and 10, also in regions not so far identified to harbor major protein QTL.

Regressions and R^2 of the EBVs of bulls with genotypes on the sum of their marker effect values are given in Table 2. All regressions were less than unity. The *y*-intercepts are also given in this table, but are to some extent arbitrary, because the directions of the marker effects estimated in the cow model were also arbitrary. The R^2 values ranged from 10% for SCS to 45% for protein production. The R^2 values tended to increase with increase in the regressions. There was no clear relationship between the R^2 values and the heritabilities. These results should be compared with the rather surprising results for human height. Despite the fact that height has a heritability of 0.8, effects from very large genome scans were able to explain only 5% of the variance (Maher, 2008). The R^2 values are also presented for the young bulls with current EBVs, but without daughter records in the truncated data set. All R^2 values for the young bulls were <0.4 of the R^2 values for all bulls. A possible explanation is that QTL, with relatively large effects that were segregating in the population in previous years, reached fixation in the group of young bulls.

The R^2 values and regression coefficients derived from the truncated data set to compute the Method II GEBVs are given in Table 4 with their standard errors. Only bulls with dam EBVs based on at least one daughter record were included. The number of bulls ranged from 673 for production traits to 741 for herd-life. The *y*-intercepts were not significantly different from zero for milk and protein ($P < 0.05$). The effect of birth year was not significant for milk, fat, persistency and herd-life. The regressions of parent average and the sum of marker effects were significant for all traits ($P < 0.001$). The regression coefficients of the SNP effects were <0.3 for all

Table 4 *Regression coefficients ± s.e. and* R^2 *used to derive the Method II GEBV*

Trait	Number of bulls	R^2	*y*-intercept	Coefficients		
				Parent average	SNP effects	Birth year
Milk (kg)	673	0.68	−1.6[ns]	0.70	0.22	−1.3[ns]
Fat (kg)	673	0.71	−9.4	0.69	0.28	−0.05[ns]
Protein (kg)	673	0.78	−1.7[ns]	0.86	0.23	−0.28
SCS	702	0.62	0.13	1.10	0.08	−0.005
Fertility (%)	706	0.74	0.37	1.11	0.07	−0.035
Persistency (%)	682	0.70	1.28	0.57	0.27	−0.014[ns]
Herd-life (days)	741	0.70	−45.9	0.76	0.29	−0.54[ns]

GEBV = genomic estimated breeding value; SNP = single nucleotide polymorphism; SCS = somatic cell score.
[1]ns, $P < 0.05$.

Table 5 *Means ± s.d. of current EBV, parent average EBV in the truncated data set and GEBV of the young bulls by trait*

Trait	Number of bulls	Reliability	Mean ± s.d.			
					GEBV	
			Current EBV	Parent average	Method I	Method II
Milk (kg)	153	0.92	−14 ± 310	102 ± 192	75 ± 418	−38 ± 253
Fat (kg)	153	0.92	1.1 ± 12.6	5.3 ± 7.8	4.0 ± 14.9	−0.6 ± 10.2
Protein (kg)	153	0.92	2.1 ± 7.6	4.5 ± 4.5	4.8 ± 9.0	0.7 ± 6.7
SCS	159	0.91	−0.01 ± 0.19	0.00 ± 0.13	0.01 ± 0.36	−0.05 ± 0.17
Fertility (%)	151	0.82	−0.68 ± 2.41	−0.66 ± 1.64	0.76 ± 4.84	−0.37 ± 2.18
Persistency (%)	145	0.90	−0.59 ± 2.25	−0.20 ± 1.51	0.61 ± 2.66	−0.40 ± 1.96
Herd-life (days)	134	0.82	−17.6 ± 90	−11.7 ± 49	7.2 ± 148.2	−28.3 ± 74

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

traits, as compared with the Method I GEBVs, which assume coefficients of unity. The R^2 values ranged between 0.62 for SCS and 0.78 for protein.

Means ± s.d. of current EBVs and GEBVs of the young bulls by trait are given in Table 5. For all three milk production traits, the parent averages were higher than the current EBVs. Thus, parent averages were biased upward. The classical explanation for this finding is preferential treatment of bull dams (e.g. Kuhn *et al.*, 1999). However, this should not be a major factor in Israel, where prices of bull calves are only slightly higher than the beef prices. One possibility is that most of the bull dams are the animals with the highest EBVs for milk production traits, and at the extreme ends of the population heritabilities are lower. Method I evaluations were also biased upward, but less so than the parent averages. Method II GEBVs were negatively biased for the milk production traits; however, in all cases the absolute value was less than for the parent averages and the Method I GEBVs. Bias for SCS was minimal for all three methods. Method I evaluations were biased upward for fertility, persistency and herd-life. Standard deviations were smallest for parent averages, as expected. Method I s.d. were the largest, and even larger than the current EBVs. Method II s.d. were between the parent averages and the current EBVs.

Correlations of current EBVs with parent averages and GEBVs of the young bulls by trait are given in Table 6.

Table 6 *Correlations of current EBV with parent average EBV and GEBV of the young bulls by trait*[1]

Trait	Parent average	GEBV	
		Method I	Method II
Milk (kg)	0.44	0.31	0.43
Fat (kg)	0.37	0.35	0.41
Protein (kg)	0.35	0.32	0.37
SCS	0.51	0.27	0.49
Fertility (%)	0.68	0.32	0.66
Persistency (%)	0.54	0.50	0.56
Herd-life (days)	0.44	0.46	0.48

EBV = estimated breeding value; GEBV = genomic estimated breeding value; SCS = somatic cell score.
[1]The numbers of young bulls by trait are given in Table 5.

The correlations with the Method I GEBVs were lower than the parent averages and Method II GEBVs for all traits. Thus, Method II is clearly superior to Method I, both by the criteria of bias and accuracy. The correlations with Method II GEBVs were higher than the correlations with parent averages for fat, protein, persistency and herd-life.

Regressions of current EBVs on parent averages, and Method II GEBVs of the young bulls by trait are given in Table 7. If the GEBVs are unbiased, regression should equal unity.

Table 7 *Regressions of current EBV on parent average EBV and Method II GEBV of the young bulls by trait*[1]

Trait	Parent average	GEBV Method II
Milk (kg)	0.71	0.53
Fat (kg)	0.60	0.51
Protein (kg)	0.59	0.42
SCS	0.74	0.55
Fertility (%)	1.00	1.21
Persistency (%)	0.81	0.64
Herd-life (days)	0.81	0.58

EBV = estimated breeding value; GEBV = genomic estimated breeding value; SCS = somatic cell score.
[1]The numbers of young bulls by trait are given in Table 5.

Table 8 *Correlations of mean daughter record deviations with parent averages, GEBV and current EBV of the young bulls by trait*[1]

		GEBV		
Trait	Parent average	Method I	Method II	Current EBV
Milk (kg)	0.39	0.26	0.37	0.97
Fat (kg)	0.21	0.25	0.29	0.95
Protein (kg)	0.26	0.27	0.31	0.95
SCS	0.34	0.21	0.34	0.92
Fertility (%)	0.34	0.25	0.36	0.79
Persistency (%)	0.32	0.28	0.42	0.91
Herd-life (days)	0.20	0.27	0.27	0.87

GEBV = genomic estimated breeding value; EBV = estimated breeding value; SCS = somatic cell score.
[1]The numbers of young bulls by trait are given in Table 5.

All regression were <1.0, except for fertility. The regressions for parent averages were higher than the Method II regressions, except for fertility.

The high parent average correlation for fertility may reflect the fact that mean reliability was lower for this trait, because of the heritability of only 2% in first parity. Therefore, the current evaluations are affected more by the parent contributions, which should increase the correlation with the parent average. The correlations of MDD with parent average, Method II GEBVs and current EBVs of the young bulls by trait are given in Table 8. Correlations of MDD with the milk production traits, SCS and persistency were between 0.91 and 0.97, but lower for fertility, which has very low heritability, and herd-life, which has only a single record per animal. Correlations of MDD with the Method II GEBVs were higher than the correlations of MDD with parent averages of all traits, except for milk and SCS, for which the two correlations were equal. The correlations of MDD with Method I were lower than the Method II correlations for all traits; however, Method I correlations were higher than the parent averages for fat, protein and herd-life.

The correlation between the sum of all marker effects from Method III analysis, Zû, and the EBVs of the bulls from the complete data set was 0.01, and was not significant. The correlation for the 186 bulls without daughter records in the truncated data set was −0.07, but also not significant. Thus, similar to the results of VanRaden *et al*. (2009a) this method was not able to generate accurate GEBVs on a data set of this size. The standard deviation among Method III evaluations was only 2.14. This is apparently because of the small assumed variance of the individual marker effects, which were assumed to sum the total additive genetic variance (VanRaden, 2008).

Discussion

Method II has the advantages as compared with the method of VanRaden (2008) in that the method of VanRaden (2008) is not effective for populations of this size, and genotypes of cows can be readily incorporated into Method II analysis without modification. Method II has the advantage over the method of Aguilar *et al*. (2010) in that computation of Method II GEBVs for new bulls with genotypes requires only calculation of a relatively simple regression equation, whereas the method of Aguilar *et al*. (2010) requires recalculation of the entire set of modified mixed model equations. An advantage of the method of Aguilar *et al*. (2010) is that the number of equations is independent of the number of markers. Thus, this method could also use all data from the high density SNP-chips with 777 000 markers.

Method II was superior to parent averages with respect to bias for production traits and persistency, and with respect to correlations with MDD for all traits, except for milk (which is not included in the Israel breeding index) and SCS. Unlike parent averages, Method II biases for production trait were negative. One of the main objections to insemination of cows with semen from young bulls with GEBVs is that, on an average, the genetic evaluations of these bulls tend to decrease when daughter records become available (e.g. http://www.altagenetics.com/English/Whatsnew/20101218Genomics.htm). This should not be the case for Method II EBVs.

Hayes *et al*. (2009) predicted that 5000 phenotypic records with genotypes should be required to obtain an accuracy of 0.6 for GEBVs with a heritability of 0.2. As heritability decreases, the number of genotypes required to reach the same accuracy increases. However, in this study, the highest correlation of Method II GEBVs with MDD was for persistency, even though this trait has moderate heritability.

VanRaden *et al*. (2009a) obtained an R^2 value of 0.47 for protein between GEBVs and daughter deviations, as compared with an R^2 of 0.27 for the parent average. Both numbers were lower in this study. Their R^2 for net merit was 0.28 for GEBVs, compared with 0.11 for parent average; however, with only 1151 predictor bulls, the comparable numbers were 0.12 for GEBVs and 0.08 for parent average. This number of predictor bulls is still greater than the one in this study. In the analysis of the Holstein, Jersey and Brown Swiss populations by VanRaden *et al*. (2009b) 4422, 1149 and 228 bulls, respectively, were included in the predictor samples. Gains in reliability for protein were 0 for Jerseys and 1% for Brown Swiss. Average gains were 11% for Jerseys, as compared with 29% for Holsteins. In 6 of the 24 traits analyzed, parent averages of

Brown Swiss were more accurate than the GEBVs. In the most recent results from the United States, regressions of June 2010 daughter deviations on August 2006 genomic evaluations ranged between 0.51 and 0.62 for milk production traits for Brown Swiss, even though 1852 bulls were genotyped (Wiggans *et al*., 2011). The R^2 values ranged from 0.21 to 0.24. Thus, the method of VanRaden (2008) is clearly inefficient for analysis of populations of the relatively moderate size analyzed in this study.

The relatively low correlation between the REML effects and the cow model effects was rather surprising. This also tends to indicate that some SNPs with potentially large effects by the cow model were excluded from the analysis. Reasons for the relatively low correlation are: first, that REML variance components for samples of this size (~1000 animals with genotypes and ~1150 ancestors) have relatively large confidence intervals; second, as SNPs on the same chromosome are correlated, the effects are reduced if multiple SNPs are included in the same analysis, relative to the situation in which each SNP is analyzed separately. Furthermore, if two SNPs are highly correlated, then attribution of the effect associated with both markers is virtually arbitrary, or dependent on allelic frequencies. Finally, in the REML analysis, each bull was weighted equally, whereas in the cow model the effects were determined based on the actual phenotypic records.

Weigel *et al*. (2009) compared GEBVs of net merit computed with all valid SNP markers to GEBVs computed with subsets of markers. The R^2 for the complete marker set was 0.375. The R^2 values for subsets of the 300, 500, 750, 1000, 1250, 1500 and 2000 SNPs with largest effects were 0.184, 0.236, 0.279, 0.289, 0.307, 0.313 and 0.322, respectively. Increasing the number of SNPs over the range of 10 000 to 40 000 increased the R^2 values generally by only 2% to 4% (VanRaden *et al*., 2009a).

Inclusion of all the markers in the method presented is not yet a viable option, because of computing limitations; however, the number of markers could be increased if the additional gain in computing time is economically justified. With respect to computing time, the limiting factor is iteration of the cow model, which required close to 24 h for completion of 2000 iterations on a HP BL860C single processor server blade. Although addition of 400 marker equations does not significantly increase the total number of equations, the cow by marker sub-matrix has no empty cells. Therefore, running time for the cow model is more than double the run time for the multitrait AM, in which the vast majority of cells in the coefficient matrix are empty.

Reducing the number of markers in the current analysis from 400 to 300 reduced R^2 of current EBVs on the sum of marker effects for protein from 0.45 to 0.24. The correlation of GEBVs with current EBVs was not reduced; however, the correlation of GEBVs with MDD was reduced from 0.31 to 0.29. Thus, a moderate increase in the number of markers included in the Method II analysis apparently will not result in a substantial improvement, at least for traits with intermediate heritability.

Conclusions

A method for marker-assisted selection based on genotype probabilities for animals that were not genotyped is presented. Correlations by Method II GEBVs were higher than the correlations with the parent averages for fat and protein production, milk production persistency and herd-life. Method II GEBVs were less biased than parent averages for milk production traits. Increasing the number of markers included in the analysis should marginally increase the accuracy of the evaluations, at least for milk production traits.

Acknowledgments

This research was supported by grants from the Israel milk marketing board, the European Sixth Research and Technological Development Framework Programme, Proposal no. 016250-2 SABRE, 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.

References

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S and Lawlor TJ 2010. *Hot topic:* a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93, 743–752.

Baruch E and Weller JI 2008. Incorporation of discrete genotype effects for multiple genes into animal model evaluations when only a small fraction of the population has been genotyped. Journal of Dairy Science 91, 4365–4371.

Baruch E and Weller JI 2009. Incorporation of genotype effects into animal model evaluations when only a small fraction of the population has been genotyped. Animal 3, 16–23.

Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Wind AE-vd, Lee J-H, 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.

Cromie AR, Berry DP, Wickham B, Kearney JF, Pena J, van Kaam JBCH, Gengler N, Szyda J, Schnyder U, Coffey M, Moster B, Hagiya K, Weller JI, Abernethy D and Spelman R 2010. International genomic co-operation; who, what, when, where, why and how? Proceedings of the Interbull Meeting, Riga, Latvia 42, 72–78.

Ducrocq V, Fritz S, Guillaume F and Boichard D 2009. French report on the use of genomic evaluation. Proceedings of the Interbull Meeting, Uppsala, Sweden, pp. 17–22.

Ezra E, Weller JI and Drori D 1987. Estimation of environmental effect of milk protein content. Heker Umas 9, 31–35 (in Hebrew).

Gianola D, de los Campos G, Hill WG, Manfredi E and Fernando R 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183, 347–363.

Goddard ME and Hayes BJ 2007. Genomic selection. Journal of Animal breeding and Genetics 124, 323–330.

Guillaume F, Fritz S, Boichard D and Druet T 2008. Short communication: correlations of marker-assisted breeding values with progeny-test breeding values for eight hundred ninety-nine French Holstein bulls. Journal of Dairy Science 91, 2520–2522.

Habier D, Fernando RL and Dekkers JCM 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177, 2389–2397.

Hayes BJ, Bowman PJ, Chamberlain AJ and Goddard ME 2009. Invited review: genomic selection in dairy cattle: progress and challenges. Journal of Dairy Science 92, 433–443.

Israel C and Weller JI 1998. Estimation of candidate gene effects in dairy cattle populations. Journal of Dairy Science 81, 1653–1662.

Kerr RJ and Kinghorn BP 1996. An efficient algorithm for segregation analysis in large populations. Journal of Animal breeding and Genetics 113, 457–469.

Kuhn MT, Freeman AE and Fernando RL 1999. Approaches investigated to correct for preferential treatment. Journal of Dairy Science 82, 181–190.

Legarra A, Aguilar I and Misztal I 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92, 4656–4663.

Liu Z, Seefried F, Reinhardt F and Reents R 2009. A simple method for correcting the bias caused by genomic pre-selection in conventional genetic evaluation. Proceedings of the Interbull Meeting, Barcelona, Spain, pp. 185–188.

Loberg A and Durr JW 2009. Interbull survey on the use of genomic information. Proceedings of the Interbull International Workshop on Genomic Information in Genetic Evaluations, Uppsala, Sweden, pp. 3–14.

Maher B 2008. Personal genomes: the case of the missing heritability. Nature 45, 18–21.

Mäntysaari E, Liu Z and VanRaden P 2010. Interbull validation test for genomic evaluations. Proceedings of the Interbull International Workshop on Genomic Information in Genetic Evaluations, Paris, France.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, OConnell J, Moore SS, Smith TPL, Sonstegard TS and Van Tassell CP 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4, e5350.

Misztal I, Legarra A and Aguilar I 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92, 4648–4655.

Party C and Ducrocq V 2009. Bias due to genomic selection. Proceedings of the Interbull Meeting 39, Uppsala, Sweden.

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

Tsuruta S, Aguilar I, Misztal I, Legarra A and Lawlor T 2010. Multiple trait genetic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, 0489_PP2-15, Leipzig, Germany.

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 2009a. Invited review: reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92, 16–24.

VanRaden PM, Wiggans GR,, Sonstegard TS and Schenkel F 2009b. Benefits from cooperation in genomics. Proceedings of the Interbull International Workshop on Genomic Information in Genetic Evaluations, Uppsala, Sweden, pp. 67–72.

Weigel KA, de los Campos G, González-Recio O, Naya H, Wu XL, Long N, Rosa GJM 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 2004. Genetic analysis of the Israeli Holstein dairy cattle population for production and non-production 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, Song JZ, Heyen DW, Lewin HA and Ron M 1998. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics 150, 1699–1706.

Weller JI, Golik M, Seroussi E, Ezra E and Ron M 2003. Population-wide analysis of a QTL affecting milk-fat production in the Israeli Holstein population. Journal of Dairy Science 86, 2219–2227.

Weller JI, Glick G, Ezra E, Zeron Y, Seroussi E and Ron M 2010. Paternity validation and estimation of genotyping error rate for the BovineSNP50 BeadChip. Animal Genetics 41, 551–553.

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.