Predictive ability of selected subsets of SNPs for moderately sized dairy cattle populations

J. I. Weller, G. Glick, A. Shirak, E. Ezra, Y. Zeron, and M. Ron

ARO, the Volcani Center

Israel Cattle Breeders Association

Sion

התאחדות מגדלי בקר בישראל

Conventional wisdom for computation of genomic EBV

1. The accuracy of genomic EBV (GEBV) computed from subsets of markers is not more than the accuracy of GEBV computed from analysis of all markers (e. g., Zhang et al., 2011, JDS).
2. The accuracy of GEBV for young bulls computed from analysis of <1000 bulls is no higher than the accuracy of parent averages (e. g., VanRaden, et al., 2009, JDS).
3. Bayesian “shrinkage” of marker effects improves accuracy of GEBV at best marginally.

Signatures of contemporary selection in the Israeli Holstein dairy cattle.

Glick et al., 2012, *Animal genetics* (In press).

…for genomic selection, monitoring the allelic frequencies in the subset of younger individuals may be more useful than monitoring allelic frequencies in the total population.

...out of the15,485 haplotypes with population frequencies between 5% and 95%, 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.

Methods proposed to select subsets of markers for analysis

- Random (Vazquez et al., 2010, JDS)
- Equally spaced throughout the genome (Habier et al., 2009, Genetics; VanRaden et al., 2009, JDS; Vazquez et al., 2010, JDS; Weigel et al., 2009, JDS; Zhang et al., 2011, JDS).
- Markers with the greatest effects as estimated from the analysis of all markers (Vazquez et al., 2010, JDS; Weigle et al., 2009, JDS; Zhang et al., 2011, JDS).

The Israeli Holstein data set and the traits analyzed

- All valid records from January, 1985, through November, 2011, were included in the analysis.
- The complete data set was divided into a “training set,” records prior to June, 2008; and the “validation set,” records from June, 2008.
- Eight traits were analyzed; milk, fat and protein production, somatic cell score (SCS), female fertility, persistency of milk production, herdlife and PD11, the Israeli breeding index.

Computation of estimated breeding values (EBV)

- Multitrait animal model EBV were computed for milk, fat, protein, SCS, female fertility, and persistency, with each parity considered a separate trait. Parities 1-5 were included in the analyses.
- Female fertility was computed as the inverse of the number of inseminations to conception.
- Single trait animal model EBV were computed for herdlife.

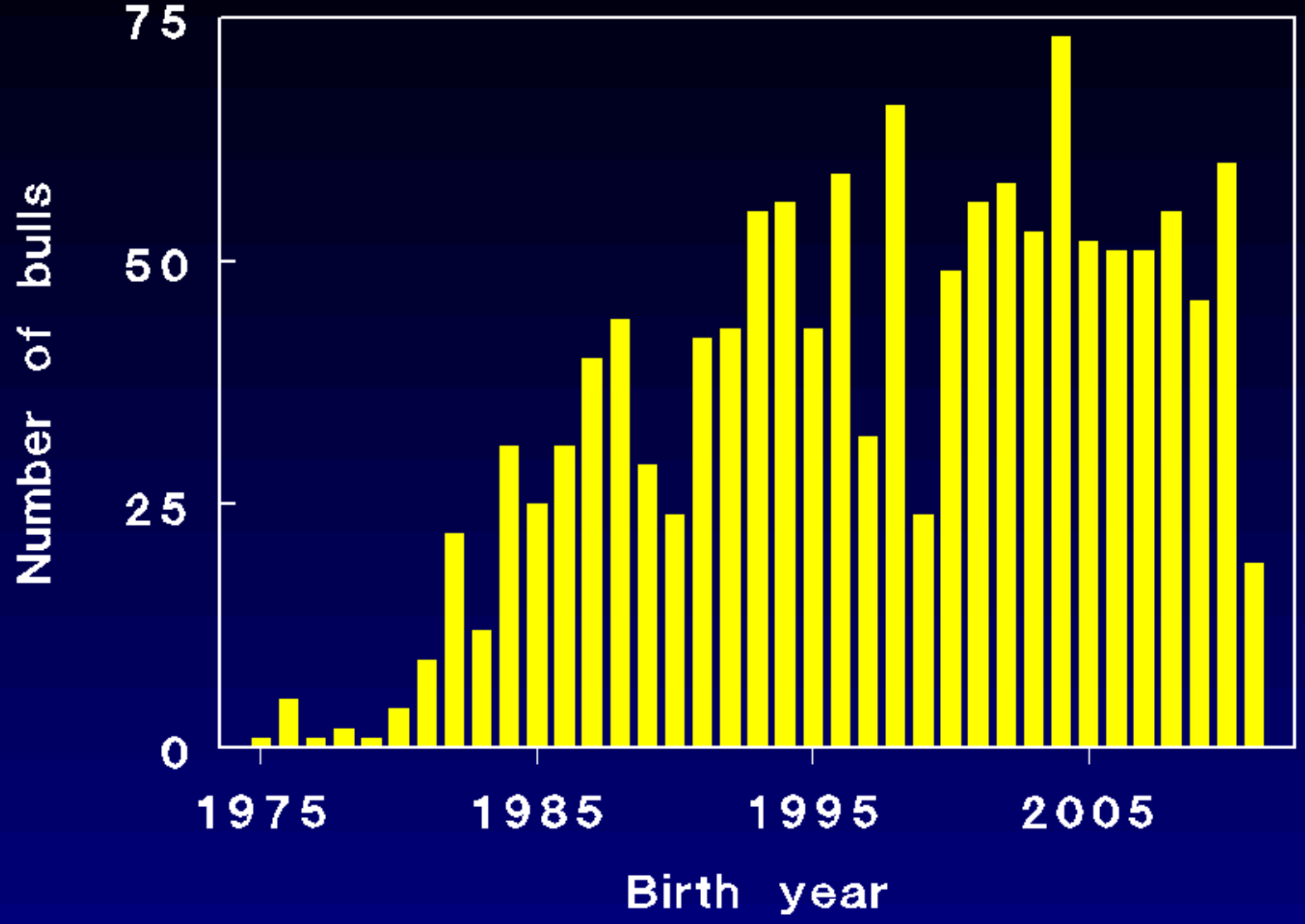
Computation of modified daughter yield deviations (MDYD)

- MDYD, weighted means of daughter records corrected for HYS 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.

Edits on genotypes

- 1359 bulls and calves were genotyped, 912 bulls for the 54,001 Bovine SNP BeadChip, and 447 for the 54,609 SNP50 v2 BeadChip.
- 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, then the second SNP was deleted.
  - After edits there were 39,816 valid SNPs.

Numbers of bulls with genotypes by birth year



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

Computation of GEBV

- The method of VanRaden (2008) was used to compute marker effects on the MDYD for each trait.
- Regression coefficients for the sum of marker effects, parent average EBV from June, 2008 (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 daughter record.

Computation and validation of GEBV

- The regression coefficients derived from the training set were then used to compute GEBV for the young bulls with MDYD in the validation set.
- GEBV were computed only for young bulls with genotypes and dam EBV based on at least one valid record.
- The GEBV of the young bulls were compared to their November, 2011, EBV and MDYD computed from the complete data set.

Computation of GEBV with all SNPs and evenly spaced SNPs

- GEBV were computed as described for protein and female fertility using all valid markers and using each 20th valid SNP.
- 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.

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. SNPs	EBV		MDYD	
		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

Conclusions from GEBV computed using all SNPs and evenly spaced SNPs

- With all valid SNPs, correlations of GEBV with current EBV and MDYD were slightly higher than PA for fertility, but lower for protein.
- With >2000 evenly spaced SNPs correlations of GEBV with current EBV were lower than PA for both traits.

Selection of markers, method 1

1. SNPs are first selected for each trait based on a linear model analysis of each marker for each trait.
2. A subset of these SNP is analyzed by a REML model including relationships. In both steps each SNP is analyzed separately.
3. The number of SNPs included in the analysis was varied over the range of 400 to 6000 to obtain an optimum.

Correlation of GEBV and current EBV as a function of number of SNPs

Correlations of Method 1 GEBV and PA with current EBV and MDYD with optimum number of SNPs

Traits	Optimum No. SNPs	Correlations			
		EBV with		MDYD with	
		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

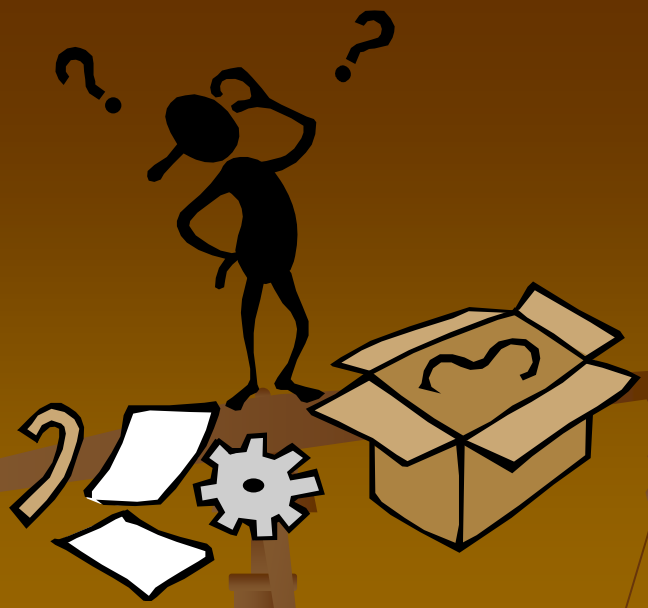
Conclusions for Method 1

- The optimum number of markers was between 600 and 6000 for all of the traits analyzed.
- The correlations of GEBV with current EBV and MDYD were higher than the correlations of PA with EBV and MDYD for all traits.
- 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.

Conclusions for Method 1

- All correlations were lower for herd-life, which has only one record per cow.
- The mean difference in the correlations between GEBV and PA was 0.18.
- For protein the correlations of EBV with GEBV and PA were 0.59 and 0.39.

The “Catch”



- In Method 1, the SNPs were selected based on their current EBV, not the EBV from the training set.
- We then selected SNPs by the same procedure, but based on their 2008 EBV (Method 2).
- **In Method 2, GEBV were not more accurate than parent averages!!!**

The rational for the difference between Methods 1 and 2

- 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, other QTL are segregating in the young bulls.
- In order to improve GEBV, it is necessary to include markers linked to QTL that are not segregating in the training set.

Method 3

- First, the SNPs with the greatest effects on the training set EBV, as determined by the REML analysis were deleted.
- Of the remaining SNPs, the markers with the greatest effects on 2011 EBV were retained.
- The numbers of SNPs deleted, based on the 2008 EBV; and retained, based on the 2011 EBV, were varied to obtain a maximum correlation between the GEBV and EBV of the young bulls.

Correlations of Method 3 GEBV and PA with current EBV and MDYD with optimum number of SNPs

Traits	Optimum No. SNPs		Correlations			
			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

Conclusions for Method 3

- 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.
- **This method also used information not available in June, 2008.**

Method 4

- First, the SNPs with the greatest effects on the training set 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.
- 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 Method 4 GEBV and parent averages with current EBV and MDYD with optimum number of SNPs

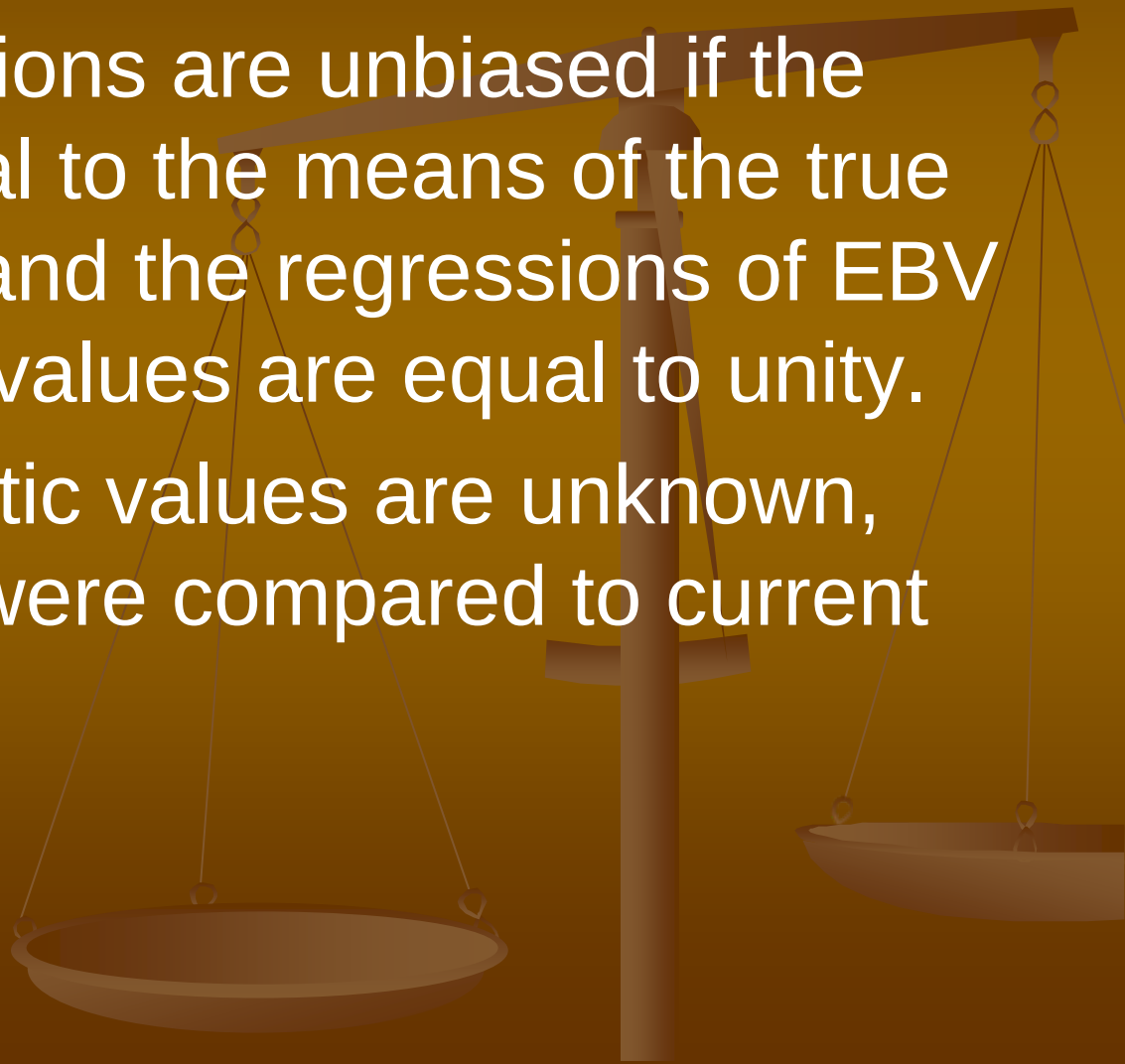
Traits	Optimum No. SNPs		Correlations			
			EBV with:		MDYD with:	
	deleted	included	PA	GEBV	PA	GEBV
Milk	500	2000	0.55	0.48	0.56	0.48
Fat	200	1000	0.44	0.52	0.34	0.43
Protein	1200	1500	0.39	0.47	0.41	0.48
SCS	1000	800	0.53	0.56	0.42	0.48
Fertility	1500	2000	0.65	0.66	0.36	0.39
Persistency	8000	9000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

Conclusions for Method 4, with respect to the correlations

- 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 did not use information not available in June, 2008!!**

Estimation of bias for Method 4 GEBV

- 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

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

Means and standard deviations of PA, 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 for Method 4 with respect to bias and coefficients of determination

- 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.
- GEBV were nearly unbiased for PD11, while the mean of PA was nearly 100 units higher than the the current EBV.
- Coefficients of determination for GEBV were higher than for PA for all traits, except for milk production

Final 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 yield GEBV that outperform parent averages, even if the training population includes <1000 bulls.
- **So much for conventional wisdom….**