Incorporation of genotype effects into Animal Model Evaluations when only a small fraction of the population has been genotyped

E. Baruch[1] & J. I. Weller[2]

[1]Faculty of Agriculture, Hebrew University of Jerusalem, Israel
[2]Institute of Animal Sciences ARO, The Volcani Center, Bet Dagan, Israel

Summary

The method of Israel and Weller (1998) to estimate QTL effects when only a small fraction of the population was genotyped was investigated by simulation. QTL effect was underestimated in all cases, but bias was greater for extreme allelic frequencies, and increased with the number of generations included in the simulations. Apparently, as the fraction of animals with inferred genotypes increases, the genotype probabilities tend to "mimic" the effect of relationships. Unbiased estimates of quantitative trait locus effects were derived by a modified "cow model" without inclusion of the relationship matrix on simulated data, even though only a small fraction of the population was genotyped. This method yielded empirically unbiased estimates for the effects of the genes *DGAT1* and *ABCG2* on milk production traits in the Israeli Holstein population. Based on these results, an efficient algorithm for marker assisted selection in dairy cattle was proposed. Quantitative trait loci effects are estimated and subtracted from the cows records. Genetic evaluations are then computed for the adjusted records. Animals are then selected based on the sum of their polygenic genetic evaluations and QTL effects. This scheme differs from a traditional dairy cattle breeding scheme in that all bull calves were considered candidates for selection. At year 10 total genetic gain was 20% greater by the proposed algorithm as compared to selection based on a standard animal model for a locus with a substitution effect of 0.5 phenotypic standard deviations. The method is flexible with respect to the model used for routine genetic evaluation. Any number of genetic markers can be easily incorporated into the algorithm, and the reduction in genetic gain due to incorrect QTL determination is minimal. It is only necessary to genotype breeding males, which are a very small fraction of the entire population.

***Key words**: Marker-assisted selection, Dairy cattle, Animal model, Quantitative trait locus*

Introduction

Most modern dairy cattle breeding programs are based on the progeny test scheme, and genetic evaluations are generally derived by application of the "animal model" (AM). Estimated breeding values (EBV) are derived for each animal, based on the production records of females and the relationship matrix including both breeding males and females. Genetic gains of ~0.1 phenotypic standard deviations per year can be obtained in modern progeny test schemes for a trait with a heritability of 0.25 (Nicholas and Smith, 1983). In trait-based selection programs only additive genetic variance is utilized, and selection is based only on animals that express the economic traits or their relatives. Trait-based selection is inefficient for low heritability traits or for traits with negative genetic correlations.

Marker-assisted selection (MAS) within a breed can increase genetic progress by increasing the accuracy of genetic evaluations, increasing the selection intensity, or decreasing the generation interval. Once a segregating quantitative trait locus (QTL) has been detected via linkage to genetic markers, application is difficult, because QTL-marker phase varies among individuals; only a small fraction of the population is genotyped; and to correctly rank candidates for selection, it is necessary to correctly weigh marker, pedigree, and trait information. Fernando and Grossman (1989) proposed a "gametic" model extension to the AM that assumes that the two QTL alleles of each individual are random effects sampled from a distribution with a known variance. Breeding values are estimated for all individuals in a population, including QTL effects via linkage to genetic markers. This method is suitable for any population structure, accommodates the fact that QTL-marker phase is generally unknown, and also can incorporate non-linked polygenic effects and other "nuisance" effects, such as herd or block. The disadvantages are that the method assumes that all animals have been genotyped, and that both recombination frequency and the variance due to the QTL are known *a priori.* Two additional equations are added to the mixed model equations for each animal for each QTL included in the analysis. Each individual with unknown parents is assumed to have two unique alleles. Thus the prediction error variances of the effects for any individual are quite large. Finally, the assumption of a normal distribution of possible QTL allele effects may not be realistic. Meuwissen and Goddard (1999) proposed methods to estimate breeding values by this approach for individuals that were not genotyped, but only for specific population structures. Recombination frequency and the variance due to the QTL can be estimated by REML for multiple linked markers, but are completely confounded for a QTL linked to a single marker (van Arendonk *et al*., 1994).

Unlike the model of Fernando and Grossman (1989), the model of Israel and Weller (1998) assumes complete linkage between the QTL and a single marker, and only two QTL alleles are segregating in the population. The model further assumes that either a daughter or granddaughter design has been applied to determine QTL genotypes of the family ancestors. The QTL effect is then included in the complete animal model analysis as a fixed effect. For individuals that are not genotyped, probabilities of receiving either allele are included as regression constants. These probabilities can be readily computed for the entire population using the segregation analysis method of Kerr and Kinghorn (1996). Israel and Weller (2002) extended this method to a situation of a QTL bracketed by two genetic markers, based on the regression analysis method of Whittaker *et al.* (1996).

The method of Israel and Weller (1998, 2002) has been tested extensively on simulated populations, and was able to yield virtually unbiased estimates of QTL effect and location, even though only 25% of the individuals were genotyped. Two and three generation populations were analyzed. However, when this model was applied to actual data from the Israeli Holstein population for the *DGAT1* locus segregating

QTL on chromosome 14 that affected milk production traits (Grisart *et al.*, 2002), the QTL effect was strongly underestimated relative to alternative estimation methods (Weller *et al.*, 2003). Reasons for this discrepancy may be due to differences between the actual and simulated data sets. The actual data set differed from the simulated data sets in three aspects. A much smaller fraction of the total population was genotyped in the actual data, <1% of the total population; frequency of one allele was very low, ~10%, in the actual data; and the actual data included ~8 generations, while the simulated data included on 2-3 generations.

The objectives of the current study were to determine the reasons for the discrepancy between the results on simulated and actual data, to modify the method of Israel and Weller (1998) so that unbiased QTL estimates are obtained, to develop a practical algorithm for MAS in which all sources of information are correctly weighted, and to test this method on simulated populations and real data.

Material and methods

Dairy cattle populations under selection for a single trait with heritability of ~0.375 were simulated. The founder population included 40 bulls and 2 000 cows. Inseminations and calvings occurred at the beginning of each year, and lactation records were available for analysis at the end of each year. The interval from insemination to calving and the interval between calvings were both assumed to be one year. Each cow could produce up to five lactations, with a probability of 0.6, for each additional lactation after first. For each cow, the probability of male or female offspring at each calving was 0.5. All female progeny became milking cows, and all male progeny were potential mating bulls. All animals were assumed to reach sexual maturity at the age of one year, and first calving of cows occurred at the age of two years.

Until year six, bulls to mate cows were selected at random among the founder bulls and all other bull calves produced that reached sexual maturity. At this point, the daughters of the founder bulls completed their second lactations, and AM genetic evaluations were computed for the first time. Based on these evaluations the five best bulls were mated to the 80 best cows to produce potential bull calves. The remaining cows were mated to the 20 best bulls including the five best. In both cases bulls were selected from those bulls selected previously as mating bulls and all one year old bull calves, including the sons of the non elite cows. Thus this scheme differs from a traditional progeny test scheme in which only bulls that have been progeny tested on a sample of test daughters are mated to the general cow population.

Bull calves not selected for breeding at the age of one year were culled. Genetic evaluations of the year-old calves were based on pedigree. In simulations which utilized QTL information, this data was also used to rank bull calves, as described below. Mating between bulls and cows were randomly determined within the elite and regular cow groups. Thus except for the five best bulls, all bulls selected for breeding would produce approximately the same number of offspring each year. After year six, genetic evaluations were computed yearly until the end of the simulation. Bulls selected for mating were not culled, and could potentially be used for mating as long as their evaluations remained among the top 20 for the general population or top 5 for elite cows.

Production records were simulated as follows:

$Y_{ijk} = a_i + p_i + h_j + q + e_{ijk}$

where:
Y_{ijk} = lactation record for cow i in herd-year-season j, of parity k;
a_i = random additive genetic effect of cow i; pi = random permanent environmental effect of cow i;
h_j = fixed effect of herd-year-season j;
q = fixed regression effect of inferred QTL genotype for cow i; and
e_{ijk} = random residual.

All effects other that q were simulated by selection from a normal distribution with a mean of zero. Variances were 0.25 for the polygenic effect, 0.125 for the permanent environmental effect, 0.25 for the herd-year-season effect and 0.5 for the residual. For the QTL we assumed that only two alleles were segregating in the population, and that the effect was codominant. Thus the variance due to the QTL = $2p(1-p)a^2$ where p = the frequency of one of the QTL alleles and a = 0.5 = the QTL substitution effect. Therefore for p = 0.5, the variance due to the QTL = 0.125; the total phenotypic variance, excluding the herd-year-season effect = 1; and heritability = 0.375. For the founders, a polygenetic effect was generated by sampling from a normal distribution with a variance of 0.25. For all other animals, polygenetic effects were generated as the sum of half the sire and half the dam genetic effects plus a "Mendelian sampling" effect generated by sampling from a normal distribution with a variance of 0.125. Thus the variance of the polygenetic effect was 0.25 for all animals. For founders QTL genotype was determined by sampling twice (for each allele) from a uniform distribution. If a value smaller than the initial allele frequency was obtained, then the individual was assumed to receive a positive allele, and otherwise the individual received the negative allele. For all other individuals QTL genotype was determined by randomly selecting one allele from each parent with a frequency of 0.5.

The first group of simulations was continued for 25 year or ~5 generations, and the total number of animals included in each simulated population was ~200 bulls and ~18 000 cows. Selection after year six was by a standard AM. These simulations were used to compare analysis models 1 and 2 described below. For estimation of QTL effects, only genotypes of bulls and one year old bull calves were assumed known. Genotypes of cows were inferred from the genotypes of their sires, as described by Israel and Weller (1998). The initial frequency of the positive QTL allele was varied from 0.1 to 0.9, but changed during the simulation due to selection on this allele. Ten simulations were computed for each value of the initial QTL allelic frequency. The second set of simulations was continued to year 30, and included ~270 bulls and ~37 000 cows. These simulations were used to compare traditional selection and the proposed MAS algorithm. Ten populations were generated for each selection scheme.

Two models for estimation of QTL effects were compared. "Model 1" was the same as Israel and Weller (1998), except that a fixed parity effect was included even though no parity effect was simulated. This effect was included because this is generally the case for analysis of commercial populations. The assumed value for the total additive genetic variance was 0.375, and the assumed values for the other variance components were equal to the simulated values. The mixed model included the inverse of the numerator relationship matrix for the polygenic effect. All ancestors, including males without records were included in computation of the relationship matrix. The assumed value for the polygenic additive genetic variance was 0.25, because the QTL was considered a fixed effect, and the assumed values for the other variance components were equal to the simulated values.

The following "cow model" was denoted "Model 2:"

$Y_{ijk} = c_i + h_j + m_k + q + e_{ijk}$

where:
c_i = random effect of cow i;
m_k = the fixed parity effect, and the other terms are as described previously.

This model differs from the model of Israel and Weller (1998) in that only cows with production records are included, and covariances among cow effects are assumed to be zero. That is the relationship matrix is not included. The assumed value for the cow variance was 0.375 (sum of additive polygenic and permanent environmental variances), because the QTL was considered a fixed effect; and the assumed values for the residual variance was 0.5, as simulated.

Although Model 2 can be used to obtain unbiased estimated of QTL effects (as will be shown), it cannot be used for routine genetic evaluation, which requires incorporation of the relationship matrix. We therefore propose the following algorithm for MAS in dairy cattle.

1. For animals with unknown genotypes for the QTL infer genotypes based on the algorithm of Kerr and Kinghorn (1996).
2. Estimate QTL effects by Model 2, as described previously.
3. Subtract the known or inferred QTL genotype effect, based on the Model 2 estimated QTL effect, from the cows' production records.
4. Compute AM breeding values for all animals from the adjusted cow records. These EBV are now based only on the polygenic effect.
5. Derive adjusted breeding values by summing the EBV with the inferred or known QTL effect of each animal.
6. Use the adjusted breeding values to rank candidates for selection.

Standard AM and "adjusted" EBV computed by the proposed algorithm were compared on the second set of simulated populations as described previously. The QTL was assumed to an initial frequency of 0.3 for the positive allele. Beginning in year 6, the simulated populations were analyzed by a standard AM and the proposed MAS algorithm was applied. In the standard breeding scheme bull calves without progeny were ranked on the mean of their sire and dam EBV. In the MAS scheme adjusted breeding values (ABV) for bull calves were computed as follows:

ABV = ½ (sire PBV) + ½ (dam PBV) + QTL effect.

Where PBV = polygenic breeding value, that is the sire and dam EBV without the QTL effects. QTL genotypes of all bull calves were assumed known. Genetic evaluations of bulls, cows, and young sires were compared to the simulated values and genetic evaluations from a standard AM. Polygenetic variance was assumed to be 0.25 in the MAS breeding scheme, and additive genetic variance was assumed to be 0.375 in the standard AM breeding scheme. Overall genetic trends and genetic trends for the polygenic and QTL effects were compared for the two selection schemes.

Results

Model 1 estimates of the QTL effects are given in Table 1. The QTL effect was underestimated in all simulations. Since there is selection for the "positive" allele, the frequencies of the two alleles in the population are closest to equality when the initial frequency of the negative allele is high. Bias is greatest when the initial frequency of the negative allele is low, because in this case the overall frequency of the negative allele is lowest, making it more difficult to accurately estimate the QTL effect. This is also reflected in the higher SD among the simulations in this

case. Bias also increased with the number of generations included in the simulations (data not shown). Thus these results conform to the results of Weller *et al.* (2003) for analysis of *DGAT1*. Bias is apparently due to confounding between the inferred QTL genotype and the relationship matrix. Apparently, as the fraction of animals with inferred genotypes increases, the genotype probabilities computed for these individuals tend to "mimic" the effect of relationships. In the extreme case of no known genotypes, the genotypes of all individuals would be inferred by the same principles used to construct the relationship matrix.

Estimates of the Model 2 QTL effect were nearly unbiased for all initial QTL frequencies. As expected, the standard deviations increased with reduction of the initial frequency of the negative allele, the allele whose frequency decreases during selection.

"Adjusted" and standard EBV for cows with records and their sires are compared in Table 2. Correlations were nearly equal to unity, but regression of adjusted EBV on the standard AM EBV were greater than unity. Evaluations of both models were nearly unbiased, that is regressions of simulated genetic values on evaluations were close to unity.

Genetic trends for total genetic gain, polygenic gain, and gain in the QTL effect from year 6 to year 24, approximately five generations of selection, are presented in Figure 1. Values prior to year 6 are not presented, because there was no genetic gain until then. The curves represent the sums of all ten simulations for each selection scheme. The trends in the simulated values are given, although the EBV trends are nearly identical. For both schemes progress is zero until year six in which the first genetic

Table 1. Effect of allele frequencies on the estimate of QTL substitution effect[1].

Initial frequency of negative allele	Estimated QTL effect	Standard deviation[2]
0.1	0.12	0.06
0.2	0.21	0.08
0.3	0.31	0.05
0.4	0.37	0.06
0.5	0.39	0.07
0.6	0.42	0.05
0.7	0.39	0.06
0.8	0.42	0.04
0.9	0.45	0.05

[1]The simulated QTL effect = 0.5.
[2]Empirical values based on the results of 10simulations for each set of initial values.

evaluations are computed. The mean annual genetic gain from year 10 through 30 was ~0.1 SD for both schemes, similar to values found in previous simulations (Nicholas and Smith, 1983). As expected, slightly greater genetic progress is obtained by traditional selection for the polygenic effect until year 15, because less selection intensity is applied to the QTL effect. The QTL and the polygenic effects can be considered a two-trait breeding objective. If greater selection pressure is applied to the QTL in the MAS scheme, then selection for the polygenic effect should be slightly less. After year 20 greater genetic progress is obtained by MAS, apparently because from this point onward the assumed variance components for the standard

AM selection scheme are no longer correct. This scheme assumed an additive genetic variance of 0.375 throughout. Greater genetic progress for the QTL effect is obtained for the MAS scheme. Since the initial frequency of the positive QTL allele is 0.3, maximum obtainable progress for the QTL is 0.7, and this value is approached for the MAS scheme by year 24. Total genetic gain was greater for the MAS scheme. The difference in genetic trends increased until year 11 and then decrease. At year 10 total genetic gain was 0.15 SD greater by MAS, but only 0.05 by year 24. Even at year 30 the MAS scheme is still higher than the traditional scheme. Thus genetic gain with trait-based selection did not surpass MAS over the long-term as noted by Gibson (1994). The "Gibson effect" was also not detected by de Koning and Weller (1994) in a series of long-term MAS simulations. The gain of 0.15 by MAS at year 10 represents an increase of nearly 20% over the traditional scheme. Alternatively, we note that the SD for milk production of Israeli Holsteins is 1 400 kg. Thus advantage of the MAS scheme = 210 kg at year 10.

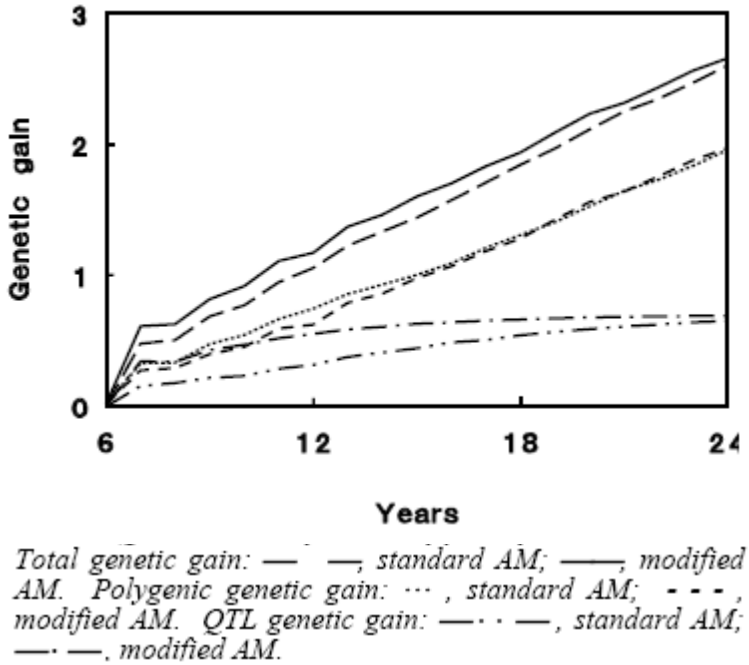

Figure 1. Total and polygenic genetic gain and genetic gain in mean QTL value as a function of year of simulation.

Table 2. Correlations between simulated genetic values and genetic evaluations, and regressions of simulated values on genetic evaluations.

	Standard AM		Adjusted model	
	Regression	Correlation	Regression	Correlation
Bulls	0.984	0.960	1.00	0.963
Cows	0.949	0.780	1.047	0.808

Conclusions

Estimates of QTL effects derived from the Model of Israel and Weller (1998) are biased due to confounding between the inferred genotypes and the relationship matrix. Unbiased estimates of QTL effects were derived by a "cow model" without relationships even though only a small fraction of the population was genotyped. An efficient algorithm for marker assisted selection in dairy cattle is presented, and tested for a QTL with a substitution effect of 0.5 phenotypic standard deviations. At year 10 total genetic gain was 20% greater by MAS as compared to standard animal model. The proposed method is easy to apply, and all required software are "on the shelf." It is only necessary to genotype males, which are a very small fraction of the entire population. The method is flexible with respect to model for general genetic evaluation, single trait AM, multitrait AM, etc. Any number of genetic markers can be easily incorporated into the algorithm.

List of references

Cohen-Zinder, M., E. Seroussi, D.M. Larkin, J.J. Loor, A. Everts-van der Wind, J.H. Lee, J.K. Drackley, M.R. Band, A.G. Hernandez, M. Shani, H.A. Lewin, J.I. Weller and M. Ron, 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.

de Koning, G.J. and J.I. Weller, 1994. Efficiency of direct selection on quantitative trait loci for a two-trait breeding objective. Theoretical and Applied Genetics 88, 669-677.

Fernando, R.L. and M. Grossman, 1989. Marker assisted selection using best linear unbiased prediction. Genetics Selection and Evolution 21, 467-477.

Gibson, J.P., 1994. Short-term gain at the expense of long-term response with selection on identified loci. In Proceedings of the Fifth World Congress for Genetics Applied to Livestock Production. Vol 21 pp. 201-204. Guelph, ON, Canada.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P., P. Simon, R. Spelman, M. Georges and R. Snell, 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research 12, 222-231.

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

Israel, C. and J.I. Weller, 2002. Estimation of quantitative trait loci effects in dairy cattle populations. Journal of Dairy Science 85, 1285-1297.

Kerr, R.J. and B.P. Kinghorn, 1996. An efficient algorithm for segregation analysis in large populations. Journal of Animal Breeding and Genetics 113, 457-469.

Meuwissen, T.H.E. and M.E. Goddard, 1999. Marker assisted estimation of breeding values when marker information is missing on many animals. Genetics Selection and Evolution 31, 375-394.

Nicholas, F.W. and C. Smith, 1983. Increased rates of genetic change in dairy cattle by embryo transfer and splitting. Animal Production 36, 341-353.

van Arendonk, J.A.M., H. Bovehuis, S. van der Beek and A.F. Groen, 1994. Detection and exploitation of markers linked to quantitative traits in farm animals.

In Proceedings of the Fifth World Congress for Genetics Applied to Livestock Production. Vol 21, pp. 193-200. Guelph, ON, Canada,

Weller, J.I., 1994. Economic Aspects of Animal Breeding. Chapman & Hall. London, UK. Weller, J.I., 2007. Marker assisted selection in dairy cattle. In: Marker-Assisted Selection, Current status and future perspectives in crops, livestock, forestry and fish. E. Guimarães, J. Ruane, B., D. Scherf, A. Sonnino and J. D. Dargie (Eds.) Food and Agriculture Organization of the United Nations. Rome, Italy, pp. 199-228.

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

Whittaker, J.C., R. Thompsom and P.M. Visscher, 1996. On the mapping of QTL by regression of phenotype on marker-type. Heredity 77, 23-32.