Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes† and M. E. Goddard†,‡

**Research Institute of Animal Science and Health, 8200 AB Lelystad, The Netherlands, †Victorian Institute of Animal Science, Attwood 3049, Victoria, Australia and ‡Institute of Land and Food Resources, University of Melbourne, Parkville 3052, Victoria, Australia*

Manuscript received August 17, 2000
Accepted for publication January 17, 2001

ABSTRACT

Recent advances in molecular genetic techniques will make dense marker maps available and genotyping many individuals for these markers feasible. Here we attempted to estimate the effects of ~50,000 marker haplotypes simultaneously from a limited number of phenotypic records. A genome of 1000 cM was simulated with a marker spacing of 1 cM. The markers surrounding every 1-cM region were combined into marker haplotypes. Due to finite population size (N_e = 100), the marker haplotypes were in linkage disequilibrium with the QTL located between the markers. Using least squares, all haplotype effects could not be estimated simultaneously. When only the biggest effects were included, they were overestimated and the accuracy of predicting genetic values of the offspring of the recorded animals was only 0.32. Best linear unbiased prediction of haplotype effects assumed equal variances associated to each 1-cM chromosomal segment, which yielded an accuracy of 0.73, although this assumption was far from true. Bayesian methods that assumed a prior distribution of the variance associated with each chromosome segment increased this accuracy to 0.85, even when the prior was not correct. It was concluded that selection on genetic values predicted from markers could substantially increase the rate of genetic gain in animals and plants, especially if combined with reproductive techniques to shorten the generation interval.

SELECTION for economically important quantitative traits in animals and plants is traditionally based on phenotypic records of the individual and its relatives. Estimated breeding values, based on this phenotypic data, are commonly calculated by best linear unbiased prediction (BLUP; Henderson 1984). One justification for molecular genetics research on livestock and crop species is the expectation that information at the DNA level will lead to faster genetic gain than that achieved based on phenotypic data only. The availability of a sparse map of genetic markers has resulted in the detection of some quantitative trait loci (QTL; Georges *et al.* 1995). The inclusion of marker information into BLUP breeding values was demonstrated by Fernando and Grossman (1989) and is predicted to yield 8–38% extra genetic gain (Meuwissen and Goddard 1996). However, the usefulness of information from a sparse marker map in outbreeding species is limited because the linkage phase between a marker and QTL must be established for every family in which the marker is to be used for selection.

The total number of single nucleotide polymorphisms (SNP) is estimated at many millions (Halushka *et al.* 1999), and the advent of DNA chip technology may make genotyping of many animals for many of these markers feasible (and perhaps even cost effective). However, the precision of mapping QTL by traditional linkage analysis is little improved by the use of a very dense marker map (Darvasi *et al.* 1993). Therefore, a different approach is needed to efficiently use all this marker information.

With a dense marker map some markers will be very close to the QTL and probably in linkage disequilibrium with it (*e.g.*, Hastbacka *et al.* 1992). Therefore, some marker alleles will be correlated with positive effects on the quantitative trait across all families and can be used for selection without the need to establish linkage phase in each family. Close markers can be combined into a haplotype. Chromosome segments that contain the same rare marker haplotypes are likely to be identical by descent (IBD) and hence carry the same QTL allele. Our approach is to estimate the effect on the quantitative trait of small chromosome segments defined by the haplotypes of marker alleles that they carry.

Quantitative traits are usually affected by many genes and consequently the benefit from marker-assisted selection is limited by the proportion of the genetic variance explained by the QTL. It would be desirable to utilize all QTL affecting the trait in marker-assisted selection. However, a dense marker map defines a very large number of chromosome segments and so there will be many effects to be estimated, probably more than there are phenotypic data points from which to estimate them.

The problem is essentially the same if we assume that

Corresponding author: Theo Meuwissen, Department of Animal Breeding and Genetics, DLO-Institute for Animal Science and Health, Box 65, 8200 AB Lelystad, The Netherlands.
E-mail: t.h.e.meuwissen@id.dlo.nl

the HUGO project and comparative mapping efforts will identify all 50,000 or so genes (*e.g.*, Aparicio 2000) in the cattle and pig genome. Hence, nearly all genes will be known, many with SNPs defined within them, and DNA chip technology will make it feasible to genotype animals for all these polymorphisms. However, when we come to estimate the allelic effects of all these genes on traits, we are again facing the estimation of very many effects in a data set of limited size, and we will not have enough degrees of freedom to fit all effects simultaneously by least squares (Lande and Thompson 1990).

In least-squares analyses, a stepwise approach can be adopted to tackle problems with insufficient degrees of freedom: genes are added to the model if they significantly improve the fit of the existing model. It seems, however, quite arbitrary to set the effects of loci to zero that are just below the significance threshold and include the full effect of those that are above this threshold. A better weighting of the information must be possible. Furthermore, selection of loci with the largest effects results in the selection of overpredicted effects. Lande and Thompson (1990) avoided this selection bias by using one-half of the data to select the loci with largest effects and using the other half to reestimate the effects. This splitting of the data set remains a suboptimal use of the information.

BLUP of allelic effects can be calculated even if there are more effects to be predicted than data points. If we assume that all loci or genes explain *a priori* an equal amount of variance (*i.e.*, the variance per locus is V_g/n, where V_g is the total genetic variance and n is the number of loci), we have only one variance to estimate. But having equal variances explained by all loci seems an unrealistic assumption.

In Bayesian statistics, parameters such as variance explained by the ith locus, V_{gi}, are assumed to come themselves from a prior distribution, $p(V_{gi})$. Hence, the variance can vary across loci, and combining of the information from the prior distribution and that of the data yields an estimate of V_{gi}. This Bayesian approach, where the variance due to each locus can vary, seems more realistic than assuming that the variance due to each locus is fixed at V_g/n, as is the case in the BLUP method.

The aim of this article is to compare least-squares, BLUP, and Bayesian analyses for their accuracy of predicting total breeding value of individuals in a situation where a limited number of recorded individuals are genotyped for many markers with many alleles per marker. Since the situation where the allelic effects of many known genes need to be estimated is very similar, we expect that the results will also hold for this situation.

METHODS

The simulated data: The alternative methods were compared by applying them to simulated data. To achieve a realistic distribution of QTL effects and gene frequencies, the simulated population was allowed to evolve until it reached an equilibrium between mutation and loss of genetic variance due to finite population size. Any particular population is likely to have experienced natural and artificial selection in the past, but this has been ignored for simplicity. The population was simulated with an effective population size of $N_e = 100$.

The genome was assumed to consist of only 10 chromosomes of 100 cM each. In the middle of each centimorgan, there was a QTL at which a mutation could occur and the QTL would become polymorphic. If the QTL is polymorphic and its allelic effects differ, it contributes to the genetic variance of the trait. At the beginning and end of every centimorgan of the chromosome a marker locus was situated; *i.e.*, there were 101 marker loci per chromosome, and every QTL was flanked by two marker loci at a distance of 0.5 cM.

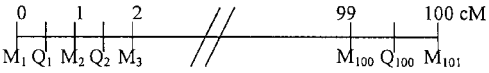
Mutations occurred randomly at the QTL at a rate of $m = 2.5 \times 10^{-5}$ per locus per generation. Since there are 1000 QTL, the mutation rate was 2.5×10^{-2} per haploid genome. For each new mutation the effect was drawn from a gamma distribution. Hayes and Goddard (2000) reviewed published estimates of QTL effects and concluded that their distribution resembled a gamma distribution with shape parameter $\beta = 0.4$. The gamma distribution with this shape was used here to simulate the effects of mutations at the QTL (see Table 1), which were assumed to be additive. Since the gamma distribution yields only positive effects, the sign of the QTL effect is sampled to be positive or negative with probability 0.5. The scale parameter of the gamma distribution was arbitrarily set to 1.66, which resulted in a genetic variance of 1 (see appendix). Because the environmental variance was also assumed to be 1, heritability was 0.5. The mutation variance added to the trait each generation was therefore $\sigma^2_m = 1000 \times m \times E(a^2) = 5 \times 10^{-3}$ environmental variance units (see appendix).

The mutation rate at the marker loci was 2.5×10^{-3} to allow a high probability of polymorphic marker loci, and every mutation at a marker locus resulted in a new unique marker allele. Hence, many marker loci were multiallelic. This might resemble the situation where several closely linked biallelic SNP markers are combined into one multiallelic marker haplotype.

To arrive at a mutation-drift balance, populations were simulated for 1000 generations at an effective size of 100. After these 1000 generations, the actual size of the populations was increased, to 200 (100 males and 100 females) in generation 1001, and to 2000 (20 half-sib families of size 100 each) in generations 1002 and 1003. The animals in generations 1001 and 1002 were marker genotyped and recorded for the trait. Phenotypic records were obtained by adding a normally distributed error term with variance 1 to the genetic value of the individuals. The 2000 animals of generation 1003 are assumed to be juveniles that did not (yet) have phenotypic records and their breeding values will be

TABLE 1

The parameters of the simulated genetic model

Map per chromosome[a]	
Number of chromosomes is the total number of morgans	10
Mutation rate of QTL	2.5×10^{-5}
Distribution of additive mutational effects	Gamma(1.66; 0.4)
Dominance of QTL effects	0
Mutation rate of marker loci	2.5×10^{-3}
Population structure	
Generations 1–1000	Ideal[b], $N = 100$
Generation 1001	Ideal[b], $N = 200$
Generation 1002	20 half-sib families, $N = 2000$
Generation 1003 and later	Ideal[b], $N = 2000$
Marker genotyping	Generations 1001 and later
Phenotypic recording	Generations 1001 and 1002

[a] M, marker position; Q, QTL position.

[b] Ideal denotes a population structure where the effective size equals the actual population size. This structure is simulated by giving every male (female) in generation $t - 1$ an equal probability of becoming the sire (dam) of animal i in generation t, which implies no selection and random mating of males and females.

estimated using marker information only. The statistical methods will be compared for their accuracy of predicting the true genetic values of the animals in generation 1003.

The data set for the estimation of the marker effects consisted of 200 and 2000 marker-genotyped and recorded animals in generations 1001 and 1002, respectively. Every animal was genotyped for 1010 (= 10 × 101) marker loci. It was assumed that the linkage phases of the tightly linked markers were known without error, but in practice they have to be estimated from the genotyping information and the family relationships between the animals. In the estimation procedures, the alleles of the marker loci that flank every centimorgan (possible QTL) of the genome are combined into one marker haplotype; *e.g.*, if the alleles of the flanking markers are 2 and 3 the haplotype allele will be denoted by 2_3. The simulation resulted in on average ~50 different haplotypes per 1-cM segment; *i.e.*, the total number of haplotype effects that needed to be estimated was ~50,000.

Least-squares estimation: Since we need to estimate 50,000 haplotype effects using 2200 phenotypic records, we cannot estimate all effects simultaneously by least squares, and some stepwise procedure for including the effects needs to be adopted. We used the following simple procedure here:

a. Perform single segment regression analyses for every segment, i, using the model

$$\boldsymbol{y} = \mu\mathbf{1}_n + \boldsymbol{X}_i\boldsymbol{g}_i + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the data vector; μ is the overall mean; $\mathbf{1}_n$ is a vector of n ones; $\boldsymbol{g}_i$ represents the genetic effects of the haplotypes at the ith 1-cM segment; $\boldsymbol{X}_i$ is the design matrix for the ith segment; and $\boldsymbol{e}$ is the error deviation. The log-likelihood of the above model is calculated as $-0.5[n \ln(\sigma_e^2) + \boldsymbol{e}'\boldsymbol{e}/\sigma_e^2]$, where n is the number of records; $\boldsymbol{e}$ and σ_e^2 denote estimates of the error deviations and error variance, respectively, with $\sigma_e^2 = \boldsymbol{e}'\boldsymbol{e}/(n - \text{Rank}([\mathbf{1}_n\ \boldsymbol{X}_i]))$. These calculations yield a log-likelihood for every segment.

b. Plot the likelihood at every segment against the position of the segment. To have a likelihood peak we need one valley to the left and one valley to the right of the peak, and we required here that log-likelihood in the valleys was at least 14 units lower than that at the likelihood peak (which was found between the left and right valley). The 14 log-likelihood units are the natural log equivalent to a LOD score of 6 units (the use of the conventional LOD score of 3 yielded too many effects for simultaneous estimation). Note that the two lower likelihoods, which are exceeded by >14, are not necessarily adjacent to the position of the likelihood peak. These likelihood peaks imply a QTL segregating at the midpoint of the chromosome segment. There are usually several likelihood peaks per chromosome.

c. Estimate the effects of the haplotypes at the QTL positions simultaneously by the model

$$\boldsymbol{y} = \mu\mathbf{1}_n + \Sigma_i\, \boldsymbol{X}_i\boldsymbol{g}_i + \boldsymbol{e},$$

where summation Σ_i is over all QTL positions corresponding to a likelihood peak and $\boldsymbol{g}_i$ was estimated at the peak. All other haplotype effects are assumed to be zero. The overall mean is also arbitrarily set to zero, because its effect cannot be distinguished from that of the fixed haplotype effects.

In this way a complete array of estimates of haplotype effects was obtained.

BLUP estimation: BLUP estimation was by the model

$$\boldsymbol{y} = \mu\mathbf{1}_n + \Sigma_i \boldsymbol{X}_i\boldsymbol{g}_i + \boldsymbol{e},$$

where summation Σ_i is over all 1000 1-cM segments. The haplotype effects are assumed random here and their variance is $E(\sigma^2_{gi}) = 0.001$ (see APPENDIX). The error variance is also needed for BLUP estimation and is set equal to its true value of 1. The estimates of $\boldsymbol{g}_i$ are obtained from the mixed-model equations (HENDERSON 1984).

Bayesian estimation: method BayesA: With Bayesian estimation the data are modeled at two levels. First there is the model at the level of the data, and second there is a model at the level of the variances of the chromosome segments. The model at the level of the data is equal to that with BLUP estimation except that the variances of the segments are $\mathrm{Var}(g_{ij}) = \sigma^2_{gi}$, which differ for every segment and are estimated by the model for the variances of the segments. The latter estimation combines the information from the prior distribution of the variances and that from the data.

The prior distribution of variances of segments was the scaled inverted chi-square distribution, $\chi^{-2}(\nu, S)$, where S is a scale parameter and ν is the number of degrees of freedom. This is a convenient choice because when the information from this prior distribution is combined with the information from the data, the resulting posterior distribution is also a scaled inverted chi square (*e.g.*, WANG *et al.* 1993),

$$\mathrm{Post}(\sigma^2_{gi}|\boldsymbol{g}_i) = \chi^{-2}(\nu + n_i, S + \boldsymbol{g}_i'\boldsymbol{g}_i),$$

where n_i is the number of haplotype effects at segment i. This posterior distribution cannot be used directly for estimation because it is conditional on the unknown $\boldsymbol{g}_i$ effects. However, Gibbs sampling is based on posterior distributions conditional on all other effects, and hence it is used for the estimation of effects and variances here. The APPENDIX shows that the mean and variance of σ^2_{gi} are 0.001 and 1.675×10^{-4}, respectively. The scaled inverted chi-square distribution with $\nu = 4.012$ and $S = 0.0020$ has the same mean and variance, and this distribution was therefore used as the prior distribution of σ^2_{gi}.

When implementing the Gibbs sampler, the variances σ^2_{gi} were sampled from the above posterior distribution. For the error variance, σ^2_e, the prior distribution was $\chi^{-2}(-2, 0)$, which yields a uniformly distributed prior, *i.e.*, a flat prior, and the conditional posterior is

$$\mathrm{Post}(\sigma^2_e|\boldsymbol{e}_i) = \chi^{-2}(n - 2, \boldsymbol{e}_i'\boldsymbol{e}_i).$$

Given the error variance and the haplotype effects, the overall mean μ is sampled from the normal distribution,

$$N[\mathbf{1}_n'\boldsymbol{y} - \mathbf{1}_n'\boldsymbol{X}\boldsymbol{g}; \sigma^2_e/n],$$

where $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2\ \boldsymbol{X}_3\ \ldots]$ is the design matrix of all haplotype effects; and $\boldsymbol{g}$ is a vector of all haplotype effects. Finally, the haplotype effects, g_{ij}, were sampled from the normal distribution,

$$N[\boldsymbol{X}_{ij}'\boldsymbol{y} - \boldsymbol{X}_{ij}'\boldsymbol{X}\boldsymbol{g}_{(ij=0)} - \boldsymbol{X}_{ij}'\mathbf{1}_n\mu; \sigma^2_e/(\boldsymbol{X}_{ij}'\boldsymbol{X}_{ij} + \lambda_i)],$$

where $\boldsymbol{X}_{ij}$ is column of $\boldsymbol{X}$ of effect g_{ij}; $\boldsymbol{g}_{(ij=0)}$ equals $\boldsymbol{g}$ except that the effect of g_{ij} is set to zero; and $\lambda_i = \sigma^2_e/\sigma^2_{gi}$.

The Gibbs sampler was run for 10,000 cycles and by graphical inspection the first 1000 cycles were discarded as burn in. The samples of $\boldsymbol{g}$ from all later cycles were averaged to obtain an estimate of the haplotype effects $\boldsymbol{g}$.

Bayesian estimation: method BayesB: In reality, the distribution of genetic variances across loci is that there are many loci with no genetic variance (not segregating) and a few with genetic variance. However, the prior density of method BayesA does not have a density peak at $\sigma^2_{gi} = 0$; in fact its probability of $\sigma^2_{gi} = 0$ is infinitesimal. Method BayesB therefore uses a prior that has a high density, π, at $\sigma^2_{gi} = 0$ and has an inverted chi-square distribution for $\sigma^2_{gi} > 0$; *i.e.*, the prior distribution is

$$\begin{aligned} \sigma^2_{gi} &= 0 && \text{with probability } \pi, \\ \sigma^2_{gi} &\sim \chi^{-2}(\nu, S) && \text{with probability } (1 - \pi), \end{aligned}$$

where $\nu = 4.234$ and $S = 0.0429$ yield the mean and variance of σ^2_{gi} given that $\sigma^2_{gi} > 0$ (see APPENDIX).

In principle the Gibbs sampling algorithm of BayesA could also be used for BayesB; however, the Gibbs sampler will not move through the entire sampling space of method BayesB. This is because the sampling of $\sigma^2_{gi} = 0$ is not possible, if $\boldsymbol{g}_i'\boldsymbol{g}_i > 0$. On the other hand, the sampling of $\boldsymbol{g}_i = 0$ has an infinitesimal probability if $\sigma^2_{gi} > 0$. This problem is resolved by sampling σ^2_{gi} and $\boldsymbol{g}_i$ simultaneously from the distribution

$$p(\sigma^2_{gi}, \boldsymbol{g}_i|\boldsymbol{y}^*) = p(\sigma^2_{gi}|\boldsymbol{y}^*) \times p(\boldsymbol{g}_i|\sigma^2_{gi}, \boldsymbol{y}^*),$$

where $\boldsymbol{y}^*$ denotes the data $\boldsymbol{y}$ corrected for the mean and all other genetic effects except $\boldsymbol{g}_i$. The above indicates that we should sample σ^2_{gi} without conditioning on $\boldsymbol{g}_i$ (in contrast to BayesA) from $p(\sigma^2_{gi}|\boldsymbol{y}^*)$, and next sample $\boldsymbol{g}_i$ conditional on σ^2_{gi} and $\boldsymbol{y}^*$ as with BayesA from $p(\boldsymbol{g}_i|\sigma^2_{gi}, \boldsymbol{y}^*)$ (note that $\boldsymbol{g}_i = 0$ if $\sigma^2_{gi} = 0$). The distribution $p(\sigma^2_{gi}|\boldsymbol{y}^*)$ cannot be expressed in the form of a known distribution and therefore Gibbs sampling cannot be applied here. We used the following Metropolis-Hastings algorithm to sample from $p(\sigma^2_{gi}|\boldsymbol{y}^*)$, where the prior distribution, $p(\sigma^2_{gi})$, is used as the driver distribution to suggest updates for the Metropolis-Hastings chain (*e.g.*, GILKS *et al.* 1996):

1. Sample $\sigma_{g(\mathrm{new})}{}^2$ from the prior distribution $p(\sigma^2_{gi})$;
2. Replace the current σ^2_{gi} by $\sigma_{g(\mathrm{new})}{}^2$ with a probability of $\mathrm{Min}[p(\boldsymbol{y}^*|\sigma_{g(\mathrm{new})}{}^2)/p(\boldsymbol{y}^*|\sigma^2_{gi}); 1]$, and go to step 1, where $p(\boldsymbol{y}^*|\sigma^2_{gi})$ denotes the likelihood of the data given variance σ^2_{gi}. Note that this likelihood equals the posterior distribution, *i.e.*, where we really want to sample from, divided by the driver/prior distribution, which is as required by the independence sampling implementation of the Metropolis-Hastings algorithm (GILKS *et al.* 1996).

The calculation of the likelihood, $p(\boldsymbol{y}^*|\sigma^2_{gi})$, is described

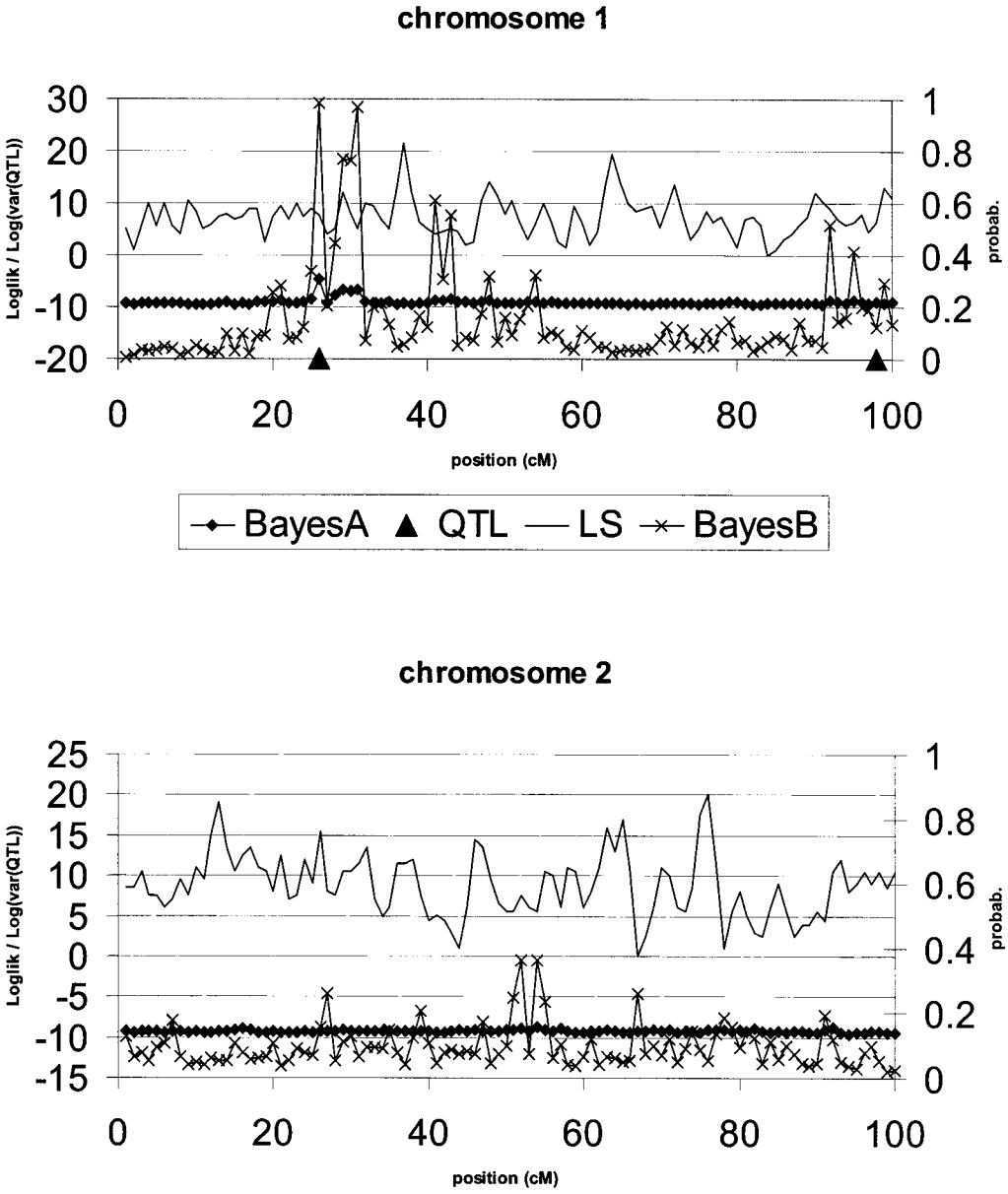


FIGURE 1.—Indicators of QTL positions at chromosomes 1 and 2 in the first replicate. BayesA, logarithm of estimated QTL variance; BayesB, posterior probability of having a QTL; LS, least-squares log-likelihood of having a single QTL as a deviation from the minimum LS-log-likelihood at the chromosome; QTL, position of the QTL that explains >2% of the total variance.

in the variance component estimation literature (*e.g.*, HENDERSON 1984). Now, the Monte Carlo Markov chain (MCMC) algorithm of BayesB consists of running a Gibbs chain as in BayesA, except that samples of σ^2_{gi} are obtained by running the above Metropolis-Hastings algorithm for 100 cycles instead of simply sampling σ^2_{gi} from an inverse chi-square distribution. The MCMC chain of BayesB was also run for 10,000 cycles and the first 1000 cycles were discarded as burn in.

Comparison of true and estimated breeding values: For all estimation methods, the estimated breeding values of the animals in generation 1003 were obtained from

$$\hat{u} = \mu \mathbf{1}_{2000} + \boldsymbol{X}_{(1003)}\boldsymbol{g},$$

where $\boldsymbol{X}_{(1003)}$ is the design matrix for generation 1003, *i.e.*, based on their marker genotyping, and $\boldsymbol{g}$ is the vector of estimates of the haplotype effects, which is obtained by least-squares, BLUP, or Bayesian estimation. Due to a recombination or a mutation, some new haplotypes occurred in generation 1003, whose effects had not been estimated in generations 1001 and 1002 and were assumed to equal zero. In the simulation, the true breeding values, $\boldsymbol{u}$, were also known in generation 1003, such that the correlation of true and estimated breeding values and the regression of true on estimated breeding values could be calculated. The latter regression coefficient equals 1 if the breeding value estimates are unbiased in the BLUP sense, *i.e.*, $E(u|\hat{u}) = \hat{u}$. The latter implies that if we select the animal with the highest $\hat{u}$, its true breeding value is expected to equal the estimate $\hat{u}$, so its breeding value is not over- or underpredicted.

RESULTS

Figure 1 shows the QTL detection results of the methods least squares (LS), BayesA, and BayesB for chromosomes 1 and 2 of one of the replicates. For LS the log-likelihoods of the individually fitted QTL are shown. For BayesA the logarithm of the estimates of σ^2_{gi} is shown, and for BayesB the posterior probability of $\sigma^2_{gi} > 0$. BLUP estimation did not provide an obvious parameter for QTL detection and was thus omitted from Figure 1. Chromosomes 1 and 2 are shown since they seemed to also characterize the results from the other chromosomes. Although there are more segregating QTL on the chromosomes, only QTL that explained >2% of the total genetic variance are shown; in fact, the QTL at positions 26 and 98 on chromosome 1 explained ~40 and 2% of the total genetic variance, respectively. Chromosome 2 was typical in that it contained no large QTL (all QTL explained <2% of genetic total genetic variance), which was the case for 5 out of 10 chromosomes in this replicate.

Although there is some correspondence between LS-likelihood peaks and the posterior probability peaks of BayesB, the LS-likelihood fluctuated much more than the posterior probability and thus yielded more false positive estimates of QTL effects (Figure 1). The method to detect QTL by LS, as described in METHODS, found QTL at positions 37 and 64 of chromosome 1 and at positions 13, 65, and 76 of chromosome 2 and found 18 QTL in total while there were 8 QTL with a $\sigma^2_{gi} > 2\%$ of the total genetic variance. The logarithm of σ^2_{gi} yielded a very flat line along most of the chromosomes and peaked only when there was a large QTL; *i.e.*, it seemed a conservative criterion for detecting QTL. The posterior probability of $\sigma^2_{gi} > 0$ of BayesB was in general a more sensitive criterion for detecting QTL, but still seemed to miss the QTL at position 98 of chromosome 1 although the posterior probability was somewhat increased in this region. In general, the really large QTL, say $\sigma^2_{gi} > 10\%$ of the total genetic variance, were detected accurately by BayesA and BayesB, but the smaller QTL were often not detected. Also, some false positive estimates were indicated by BayesB, although usually some smaller QTL were in the neighborhood of the indicated position, which may together with sampling error have increased the posterior probability of having a QTL.

When the estimates of the haplotype effects were used to estimate the true breeding values (TBV) of the animals in generation 1003, the correlations between true and estimated breeding values (EBV) were as given in Table 2. Because method BayesA required much more computer time and had poorer results than BayesB, it was investigated only in replicate 1. For the methods LS, BLUP, and BayesB, replicate 1 yielded correlations and regressions that were very similar to the averages shown in Table 2, except that BLUP yielded only $b_{TBV;EBV} = 0.844$ in this replicate.

TABLE 2

Comparing estimated *vs.* true breeding values in generation 1003

	$r_{TBV;EBV}$ + SE	$b_{TBV;EBV}$ + SE
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
BayesA	0.798	0.827
BayesB	0.848 + 0.012	0.946 + 0.018

Mean of five replicated simulations, except for BayesA which is based on one replicate. LS, least squares; BLUP, best linear unbiased prediction; BayesA, Bayesian method with inverse chi-square prior distribution; BayesB, Bayesian method where the prior density of having zero QTL effects was increased; $r_{TBV;EBV}$, correlation between estimated and true breeding values (equals accuracy of selection); $b_{TBV;EBV}$, regression of true on estimated breeding value.

The correlations of Table 2 reflect the accuracy of selection when selection is for the marker-based breeding value estimates. The accuracy of selection for LS predictions of TBV was rather low, which is probably due to the poor detection of QTL by LS (Figure 1), and because the estimation of the allelic effects of on average 15.4 detected marker haplotypes resulted in on average 872 equations. Hence, there were only 2200/872 = 2.5 d.f. per estimated effect, which would have resulted in large sampling errors and thus poor predictions of TBV. BLUP resulted in a reasonably high accuracy of selection, despite incorrectly assuming equal σ^2_{gi} for all loci. BayesA resulted in ~9% more accuracy than BLUP and the accuracy of BayesB exceeded that of BLUP by ~16%. The selection accuracies of BLUP, BayesA, and BayesB are very high for schemes where the animals have no performance information of their own. For comparison, a pedigree-based selection index would result in an accuracy of selection of ~0.4.

Table 2 also shows the regression of true breeding values on estimated breeding values. This regression should be 1 for methods that are unbiased in the BLUP sense (see METHODS). This regression coefficient was substantially <1 for LS, which indicates that the EBV would need to be regressed back to their mean to become an unbiased predictor of the TBV; *i.e.*, the estimated breeding values are too variable. In fact the LS-EBV were substantially more variable than the TBV (result not shown). The regression coefficient became closer to 1 if BLUP, BayesA, and, especially, BayesB were used, but was still somewhat <1. The small deviation of $b_{TBV;EBV}$ from 1 that remained may occur because the inverted chi-square distribution, which was used as a prior, is not equal to the simulated distribution of variances at the segregating loci, which is due to differences between gamma-distributed effects of mutations.

Table 3 shows correlations between TBV and EBV and regression coefficients of TBV on EBV when the estimation of marker haplotype effects was based on a

TABLE 3

Correlations between true and estimated breeding values when the number of phenotypic records is varied

	No. of phenotypic records		
	500	1000	2200
LS	0.124	0.204	0.318
BLUP	0.579	0.659	0.732
BayesB	0.708	0.787	0.848

reduced number of phenotypic records. The situation with 1000 or 500 records was obtained by deleting 1200 and 1700 records, respectively, of those animals of generation 1002, which had the smallest number of offspring in generation 1003; *i.e.*, the animals with most offspring were recorded. Table 3 does not show the results for method BayesA because of its computational costs and because it is expected to yield lower correlations than BayesB in any case. When the number of records was reduced from 2200 to 500, the correlation between TBV and EBV was reduced by 61, 21, and 17% for methods LS, BLUP, and BayesB, respectively. As expected, LS is least able to handle situations with few records and many effects to estimate and therefore showed a much larger reduction of the correlation between TBV and EBV as the number of records decreased. Method BayesB maintained a reasonably high correlation of 0.708 even when the estimation of the haplotype effects was based on only 500 phenotypic records.

Table 4 investigates the effects of having a less dense marker map, *i.e.*, where markers are spaced at every 2 or 4 cM. The situation with a marker distance of 2 cM was obtained by omitting every second marker from the original data set, and the situation with a 4-cM distance was obtained by again omitting every second marker. Note that the number of possible QTL positions remained at 100 per chromosome. For the BLUP analysis, σ^2_{gi} was increased from 0.0028 to 0.0056 and 0.0112, respectively. For BayesB the prior probability of having a QTL, $(1 - \pi)$, was increased from 0.053 to 0.106 and 0.212, respectively. For LS, the correlation between EBV and TBV increased as the distance between the markers increased. This is probably because the number of effects that are to be estimated is reduced when the number of markers is reduced, and the estimation of fewer effects reduces the shortage of degrees of freedom for LS estimation. BLUP and BayesB showed a small reduction in accuracy of ~4% when the marker spacing increased from 1 to 2 cM, and a 9–13% lower correlation when the spacing increased to 4 cM. This suggests that for the current population structure with $N_e = 100$, a marker spacing of 2 cM still yields sufficiently large linkage disequilibria between markers and QTL to predict the QTL effects. With the larger marker spacing of 4 cM, these linkage disequilibria are reduced.

TABLE 4

Correlations between true and estimated breeding values when the density of the marker map is varied and effective population size is 100

	Marker spacing (cM)		
	1	2	4
LS	0.318	0.354	0.363
BLUP	0.732	0.708	0.668
BayesB	0.848	0.810	0.737

DISCUSSION

Methods were presented for the estimation of allelic effects of marker or gene loci. The methods were compared in a situation where the allelic effects of small marker haplotypes surrounding 1-cM regions had to be estimated. Figure 1 indicated that BayesA and BayesB were able to predict the position of large QTL (say $\sigma^2_{gi} >$ 10% of the total genetic variance), but often did not identify the smaller QTL. However, the posterior probabilities of these smaller QTL were low but not zero for any of the possible QTL positions, so the small QTL still contributed to the prediction of total genetic values. The contributions of small QTL in the Bayesian methods were probably similar to their contribution in the BLUP prediction of total genetic values, where small and equal σ^2_{gi} are used for all loci. Because BLUP resulted in a reasonably high accuracy of predicting TBV, it seems that a correct positioning of QTL is not essential to achieve this. However, BayesA and especially BayesB did identify the positions of the largest QTL, which probably contributed to its increased accuracy of predicting TBV over BLUP.

The markers used in the simulations more closely resembled microsatellite markers than SNPs, which are biallelic and have a much lower mutation rate. However, three to five closely linked biallelic SNP markers may be pooled to obtain ~2^3 different haplotypes, which resemble the about seven alleles per marker that were used in the simulation. If the closely linked markers are within a region of ~0.25 cM, their recombination rate would resemble the mutation rate of the simulated markers (Table 1). The construction of haplotypes from the SNP markers, however, requires knowledge about the linkage phase of the markers. This requires at least two generations of typed individuals (*i.e.*, generations 1001 and 1002 here) and should be possible with high precision when the markers are closely linked; *i.e.*, (double) recombinations are very unlikely. In situations where the linkage phases are still uncertain, this uncertainty may be accounted for in the design matrix of the haplotype, $\boldsymbol{X}_i$, by having p and $(1 - p)$ at the elements

that belong to haplotypes A and B, instead of 0 and 1 (or 1 and 0). The information contributed by this record to estimate the difference between haplotypes A and B will be reduced, but this record still contributes to the differences between A (B) and all other haplotypes, such that the overall loss of information may be limited.

The accuracy with which breeding values can be predicted from the markers is limited by two factors. First, the linkage disequilibria between markers and QTL may be incomplete so that the marker haplotypes do not explain all the variance at the QTL. The results in Table 4 suggest that, in a population with $N_e = 100$, a marker spacing of 2 cM yields almost a maximum accuracy of prediction of the QTL effects. Because linkage disequilibria are a function of $N_e c$ (Sved 1971; Goddard 1991), Table 4 may more generally be interpreted as investigating the effect of $N_e c$ on accuracy of prediction of TBV; *i.e.*, the term "marker distance" could be replaced by $N_e c$, where c denotes distance between markers (in morgans). It follows from Table 4 that $N_e c$ should be <2 ($= 100 \times 0.02$) to achieve close to maximum accuracy of prediction of TBV given the information content of the current markers (the heterozygosity of markers was ~50%). With more informative markers, larger values of $N_e c$ may be used. Also, if estimation of the haplotype effects is restricted to a part of the population (*e.g.*, a few related families, the elites of a breeding scheme), the linkage disequilibrium within such a group may be much larger than expected based on $N_e c$ (Farnir *et al.* 2001) and prediction errors will be reduced within this part of the population. However, these estimates of haplotype effects may not be very useful to predict breeding values outside this part of the population.

The second effect that limits the accuracies of selection in Table 2 is the sampling error on the estimates of the haplotype effects. These sampling errors increase if the environmental variance divided by the number of genotyped and recorded animals increases. Hence, with a reduced heritability (h^2) such that ($1 - h^2$) becomes twice as large, the number of records needs to be doubled to achieve a similar accuracy. Table 3 shows that when the number of records increases from 1000 to 2200 the accuracy of predicting TBV still increases substantially. It seems that in these situations where very many effects are estimated from a limited number of records, a doubling of the number of records will yield, in practical situations, a substantial increase in accuracy (even when the number of records was already large). When haplotype effects were estimated from only 500 records, however, the accuracy of predicting TBV using BayesB was still much higher than could be expected from a pedigree-based selection index.

It may be expected that when going from a 10-M to a 30-M genome, the prediction of effects of individual chromosome segments becomes poorer because three times as many effects need to be predicted. However, in this case the EBV of an animal equals the sum of three times as many segments, such that prediction errors of individual segments are averaged out over three times as many effects. Hence, it may be expected that accuracies of selection with a 30-M genome will be similar to those of Table 2. When going from a 1-M genome (in preliminary tests of the programs) to 10-M genomes, we also found very similar accuracies of selection (although the 1-M results were more variable because sometimes there was very little genetic variance to be predicted).

The four methods of analysis used increasingly informative prior distributions for the σ^2_{gi}. LS performs badly because it greatly overestimates some haplotype effects and underestimates others. BLUP, although it uses a very simple prior, regresses estimates back toward zero, especially if there are few individuals carrying a particular haplotype. In the case of LS, better model selection methods can be used to determine which QTL effects should be included in the model; *e.g.*, start with the largest QTL and next include the second largest and so on until the QTL become too small to be included. However, the overestimation problems remain and in view of the huge number of possible QTL models and the poor results of LS in Table 2, it seems that more sophisticated methods are needed such as the use of prior distributions.

The BLUP method that was used here could also be improved upon. First, the total genetic variance could have been estimated by REML (residual maximum likelihood; Patterson and Thompson 1971) within a replicate and dividing the estimated genetic variance over the segments instead of the expected genetic variance (although, when the genome is sufficiently large, there will be little difference between these two values). The main problem with BLUP, however, remains, namely the really big QTL will be too heavily regressed back to zero. Second, the variances of the segments of QTL with large effects may be estimated by REML for use in the BLUP analysis. This leads to model selection problems as in the case of LS (which segments should be included in the analysis), and furthermore the estimation of many variances of chromosomal segments may be computationally as demanding as the estimation of these variances by BayesB.

It was assumed here that gene effects are additive, while some degree of dominance will probably occur in practice. In the presented models, only additive effects were fitted such that only the "average effects" of the genes (Falconer and Mackay 1996) are estimated, which is appropriate for the prediction of breeding values. This is probably satisfactory in many situations, except when prediction of dominant gene actions is important. In the latter case, dominance effects may be included in the model. Also, it was assumed here that to obtain the prior distribution of σ^2_{gi} mutation rate and the distribution of mutational effects were known. Although these parameters can be estimated in a metaanalysis (Hayes and Goddard 2001), they may still be dif-

ferent in specific situations. For example, the selection method may be different from that used in the populations of the metaanalysis, or the trait may be closely related to fitness, which renders most mutational effects negative. However, because the differences between BLUP, BayesA, and BayesB are relatively small (Table 2), the effect of using an incorrect prior distribution of σ^2_{gi} on the accuracy of selection seems to be small. The results of Table 2 indicate that (a) we need a proper prior distribution to avoid the overestimation problems of LS; and (b) the prior distribution should allow for small (or no) QTL effects with a high probability and for large QTL effects with a low probability, and the exact shape of the distribution seems of lesser importance (also since the inverted chi-square prior of BayesB did not perfectly agree with the gamma distribution of the QTL effects).

The simulation model assumes that every centimorgan contains a QTL that might affect the trait if a mutation occurred. However, in practice, only ~5% of these potential QTL were segregating. Thus the statistical model can be described as allowing that each marker bracket could contain a QTL. In reality, the QTL that can affect the trait will show some distribution across the genome and might be found in clusters close to each other; *i.e.*, some segments of 1 cM can contain more than one QTL. However, the effects of several closely linked QTL may be reasonably well approximated by one QTL with an increased genetic variance. Furthermore, the gamma distribution of QTL was based on QTL detection experiments (Hayes and Goddard 2001), which detect the effects of chromosomal segments rather than individual genes. In this view, the distribution of the effects of the 1-cM chromosomal segments may have been reasonably well reflected by the gamma distribution that was used here.

The main computation problem of the methods presented is that very many effects are fitted simultaneously, here 1000, such that the information matrix ($\boldsymbol{X'X}$) cannot be stored in the RAM memory of the computer. Therefore, the models were solved by the iteration on the data technique (Schaeffer and Kennedy 1986). This iterative technique, however, requires that, after the solutions of one haplotype effect are updated, the right-hand sides of all other effects are adjusted for the new solutions; *i.e.*, computer time increases approximately quadratically with the number of effects fitted. Method BayesA was most computer intensive and took ~2 weeks on a Pentium500 PC. Method BayesB required much less CPU, *i.e.*, ~1 day, because many effects have $\sigma^2_{gi} = 0$ in any cycle of the chain and thus do not enter the equations. In conclusion, the presented methods are computer intensive but are feasible on large computers even for genome sizes of 30 M or more.

The accuracies of selection in Table 2 are comparable to those obtained after a progeny test and are very high for animals without performance or progeny records.

TABLE 5

The correlation between estimated and true breeding values in generations 1003–1008, where the estimated breeding values are obtained from the BayesB marker estimates in generations 1001 and 1002

Generation	$r_{TBV;EBV}$
1003	0.848
1004	0.804
1005	0.768
1006	0.758
1007	0.734
1008	0.718

The generations 1004–1008 are obtained in the same way as 1003 from their parental generations.

Their use would dramatically increase the rate of genetic gain especially in traits where selection on phenotypic records is difficult, such as traits displayed only in females or after slaughter, disease resistance traits, or traits with low heritability. Further improvement could be had by combining high accuracies of selection with very short generation intervals to increase the number of selection cycles per unit of time. Georges and Massey (1991) and Haley and Visscher (1998) took this idea to the extreme in their "velogenetics" schemes for cattle, where oocytes were harvested from *in utero* calves (or obtained from *in vitro* meiosis of cultured cells), matured *in vitro*, fertilized, selected on the basis of their marker genotypes, and implanted in recipient cows (or cultured again), resulting in generation intervals of 6 months or less. Such a process could be repeated for several "generations" using method BayesB to predict the breeding values of the fertilized oocytes.

In velogenetics schemes the decline of the accuracy of selection over generations determines how often the haplotype effects need to be reestimated. This decline of accuracy is ~5% per generation between generations 1003 and 1005 (Table 5), and becomes smaller in later generations. This reduction is much larger than expected based on the recombination rate between a QTL and its nearest markers, *i.e.*, 0.5%. The latter indicates that more distant markers also contributed to the high accuracy of prediction of TBV. Or, putting it a different way, BayesB does not accurately predict the genetic value of individual 1-cM chromosome segments; instead it accurately predicts the total genetic value of larger chromosome segments (of, say, 4 cM or more), and therefore its accuracy reduces markedly as the large chromosome segments break up due to several generations of recombination. A larger effective population size will decrease the size of IBD chromosomal segments and will therefore improve the prediction of genetic value of small chromosome segments, provided that a sufficient number of phenotypic records is available. However, the accuracy of prediction of TBV in genera-

tion 1008, *i.e.*, six generations after the estimation of the haplotype effects, is still 0.718, which seems sufficiently high to make velogenetics breeding schemes successful. However, selection will cause changes of haplotype frequencies, which will make some haplotypes more important that were previously at low frequency, and thus less accurately estimated. Also, nonadditive genetic effects would cause a larger reduction in accuracy during generations 1003–1008 than shown in Table 5.

This study considered the selection of animals without phenotypic records. However, the results also apply to cases of genetic counseling, where the genetic risk of individuals for multifactorial diseases, *i.e.*, diseases that can (in part) be caused by many different mutations at different loci, is assessed. Alternatively, the response of individuals to various treatments can be predicted on the basis of marker information in situations where the responses to treatments depend on the genetic disposition of the individuals. Furthermore, well-chosen informative prior distributions in Bayesian estimation methods may prove useful in other examples of data mining where there are too few degrees of freedom for conventional statistical models.

CONCLUSIONS

1. By using a dense marker map covering all chromosomes, it is possible to accurately estimate the breeding value of animals that have no phenotypic record of their own and no progeny.
2. This requires the estimation of a large number of marker haplotype effects. Using least squares, all haplotype effects could not be estimated simultaneously. Even when only the largest effects were included, they were overestimated and the accuracy of predicting breeding value was low.
3. Methods that assumed a prior distribution for the variance associated with each chromosome segment gave more accurate predictions of breeding values even when the prior was not correct.
4. Selection on breeding values predicted from markers could substantially increase the rate of genetic gain in animals and plants especially if combined with reproductive techniques to shorten the generation interval.

LITERATURE CITED

Aparicio, S. A. R. J., 2000 How to count human genes. Nat. Genet. **25:** 129–130.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetic Theory.* Harper & Row, New York.

Darvasi, A., A. Weinreb, V. Minke, J. I. Weller and M. Soller, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134:** 943–951.

Falconer, D. S., and T. F. S. Mackay, 1996 *An Introduction to Quantitative Genetics.* Longman Group, Essex, UK.

Farnir, F., W. Coppieters, J.-J. Arranz, P. Berzi, N. Cambisano *et al.*, 2000 Extensive genome-wide linkage disequilibrium in cattle. Genome Res. **10:** 220–227.

Fernando, R. L., and M. Grossman, 1989 Marker-assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21:** 246–477.

Georges, M., and J. M. Massey, 1991 Velogenetics, or the synergistic use of marker assisted selection and germ-line manipulation. Theriogenology **35:** 151–159.

Georges, M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto *et al.*, 1995 Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics **139:** 907–920.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter, 1996 *Markov Chain Monte Carlo in Practice.* Chapman & Hall, New York.

Goddard, M. E., 1991 Mapping genes for quantitative traits using linkage disequilibrium. Genet. Sel. Evol. **23** (Suppl. 1): 131s–134s.

Haley, C. S., and P. M. Visscher, 1998 Strategies to utilize marker—quantitative trait loci associations. J. Dairy Sci. **81**(2): 85–97.

Halushka, M. K., J.-B. Fan, K. Bentley, L. Hsie, N. Shen *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239–247.

Hastbacka, J., A. De La Chapelle, I. Kaitila, P. Sistonen, A. Waever *et al.*, 1992 Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat. Genet. **2:** 204–211.

Hayes, B. J., and M. E. Goddard, 2001 The distribution of effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. **33**(3).

Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding.* University of Guelph, Guelph, Ontario, Canada.

Lande, R., and R. Thompson, 1990 Efficiency of marker assisted selection in the improvement of quantitative traits. Genetics **124:** 743–756.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

Meuwissen, T. H. E., and M. E. Goddard, 1996 The use of marker haplotypes in animal breeding schemes. Genet. Sel. Evol. **28:** 161–176.

Patterson, H. D., and R. Thompson, 1971 Recovery of inter-block information when block sizes are equal. Biometrika **58:** 545–554.

Schaeffer, L. R., and B. W. Kennedy, 1986 Computing solutions to mixed model equations. Proc. 3rd World Congr. Genet. Appl. Livest. Prod. **12:** 382–389.

Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. **2:** 125–141.

Wang, C. S., J. J. Rutledge and D. Gianola, 1993 Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet. Sel. Evol. **25:** 41–62.

Communicating editor: C. Haley

APPENDIX

Expectation and variance of genetic variance due to QTL in a mutation-drift model: The genetic variance due to an additive locus equals

$$\sigma^2_{gi} = 2p(1 - p)a^2$$

(Falconer and Mackay 1996), where $2p(1 - p)$ represents the heterozygosity, H, and a = one-half the difference between the performances of the two homozygotes. Because H and a^2 are independent,

$$E(\sigma^2_{gi}) = E(H) \times E(a^2).$$

$E(a^2) = E[(0.5(a_1 - a_2))^2] = 0.5\ E(a_1^2)$, where a_1 (a_2) = effect of homozygote carrying the first (second) mutation, and the above is because $E(a_1^2) = E(a_2^2)$ and $E(a_1 a_2) = 0$. The effect a_1 was sampled from the

gamma(1.66; 0.4) distribution, where the shape parameter (0.4) was estimated by Hayes and Goddard (2001) and the scale parameter is chosen such that the total genetic variance equals 1 (see next paragraph). The second moment of this gamma distribution is $E(a_1^2) = 0.203$, *i.e.*, $E(a^2) = 0.102$.

The expected heterozygosity of QTL is

$$E(H) = 4N_e m/(4N_e m + 1) = 0.0099$$

(Lynch and Walsh 1998), where m is the mutation rate (2.5×10^{-5}) per haploid locus per generation, and N_e is the effective population size (100). Hence, $E(\sigma_{gi}^2) = 0.102 \times 0.0099 = 0.001$. So 1000 QTL result in an expected total genetic variance of 1. Thus, an environmental variance of 1 yields a heritability of 0.5. The variance generated by new mutations each generation equals $2 \times 1000 \times m \times E(a^2) = 5 \times 10^{-3}$ environmental variance units.

The above E(H) is unconditional on whether the locus is segregating or not, as required for method BayesA. Method BayesB requires the calculation of E(H) given that the locus is segregating. Since $4N_e m \ll 1$, the allele frequency distribution of a segregating locus is U-shaped and approximated by

$$f(p) = K/[p(1-p)]$$

(Crow and Kimura 1970), where p = allele frequency, and the constant $K = 0.5/\ln(2N_e - 1)$. Using this distribution, the expected heterozygosity is

$$E(H|s=1) = \int_0^1 2p(1-p)f(p)\,dp = 1/\ln(2N_e - 1) = 0.1889,$$

where $s = 1$ indicates that the locus is segregating. Hence, $E(\sigma_{gi}^2|s=1) = 0.1889 \times 0.102 = 0.019$.

The prior distribution of method BayesB also involves the probability that the QTL is not segregating, π. Because $E(H) = E(H|s=1) \times \text{Prob}(s=1)$, where $\text{Prob}(s=1) = 1 - \pi$, we have

$$\pi = 1 - E(H)/E(H|s=1) = 1 - 0.0099/0.1889 = 0.947.$$

Next we need the variance of $(\sigma_{gi}^2|s=1)$, which is approximated by

$$V(\sigma_{gi}^2|s=1) = [E(a^2)]^2 V(H|s=1) + [E(H|s=1)]^2 V(a^2).$$

$V(a^2)$ is obtained from the gamma(1.66, 0.4) distribution and equals 0.08. Further,

$$V(H|s=1) = E(H^2|s=1) - [E(H|s=1)]^2,$$

where $E(H^2|s=1) = 2K/3$, which is obtained by a similar integration as that of $E(H|s=1)$. This gives $V(H|s=1) = E(H|s=1)[1/3 - E(H|s=1)] = 0.0273$. Substitution of these terms in the above equation for $V(\sigma_{gi}^2|s=1)$ yields $V(\sigma_{gi}^2|s=1) = (0.102)^2 \times 0.0273 + (0.1889)^2 \times 0.08 = 0.00315$.

Finally we need the variance of σ_{gi}^2 unconditional on the segregation status of the QTL,

$$V(\sigma_{gi}^2) = E_s[V(\sigma_{gi}^2|s)] + V_s[E(\sigma_{gi}^2|s)],$$

where $E_s[]$ ($V_s[]$) denotes taking expectation (variance) over the segregation status, s. Because $V(\sigma_{gi}^2|s=0) = 0$ if the locus is not segregating, we have

$$E_s[V(\sigma_{gi}^2|s)] = (1-\pi)\,V(\sigma_{gi}^2|s=1) = 0.0001670,$$

where $(1 - \pi)$ is the probability that $s = 1$. Since the segregation status follows a binomial distribution with probability π yielding $E(\sigma_{gi}^2|s=0) = 0$, and probability $(1 - \pi)$ yielding $E(\sigma_{gi}^2|s=1)$, we have

$$V_s[E(\sigma_{gi}^2|s)] = \pi(1-\pi)\,[E(\sigma_{gi}^2|s=1)]^2 = 4.98 \times 10^{-7}.$$

Hence, the unconditional variance of σ_{gi}^2 is $V(\sigma_{gi}^2) = 0.0001670 + 4.98 \times 10^{-7} = 0.0001675$.

The inverted chi-square distribution as prior distribution: The aim here is to find an inverted chi-square distribution $x \sim \chi^{-2}(\nu, S)$ with parameters ν and S such that the mean and variance of x equal that of σ_{gi}^2 (BayesA) or $\sigma_{gi}^2|s=1$ (BayesB). From the inverted chi-square distribution $E(x) = S/(\nu - 2)$ and $[\text{CV}(x)]^2 = V(x)/[E(x)]^2 = 2/(\nu - 4)$, where CV( ) denotes coefficient of variance. Substituting the mean and variances of σ_{gi}^2 and $\sigma_{gi}^2|s=1$ that were obtained in the previous section of the appendix and backsolving for ν and S yields the prior distribution $\sigma_{gi}^2 \sim \chi^{-2}(4.012; 0.0020)$, which was used for method BayesA, and the prior distribution $\sigma_{gi}^2|s=1 \sim \chi^{-2}(4.2339; 0.0429)$, which was used for method BayesB.