Genomic selection using different marker types and densities

Genomic selection using different marker types and densities

T. R. Solberg, A. K. Sonesson, J. A. Woolliams and T. H. E. Meuwissen

J ANIM SCI 2008, 86:2447-2454.
doi: 10.2527/jas.2007-0010 originally published online April 11, 2008

Genomic selection using different marker types and densities

T. R. Solberg,*[1] A. K. Sonesson,† J. A. Woolliams,*‡ and T. H. E. Meuwissen*

*University of Life Sciences, Department of Animal and Aquacultural Sciences, PO Box 5003, N-1432 Ås, Norway; †AKVAFORSK (Institute of Aquaculture Research Ltd.), PO Box 5010, N-1432 Ås, Norway; and ‡Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, United Kingdom

ABSTRACT: With the availability of high-density marker maps and cost-effective genotyping, genomic selection methods may provide faster genetic gain than can be achieved by current selection methods based on phenotypes and the pedigree. Here we investigate some of the factors driving the accuracy of genomic selection, namely marker density and marker type (i.e., microsatellite and SNP markers), and the use of marker haplotypes versus marker genotypes alone. Different densities were tested with marker densities equivalent to 2, 1, 0.5, and $0.25N_e$ markers/morgan using microsatellites and 8, 4, 2, and $1N_e$ markers/morgan using SNP, where $1N_e$ markers/morgan means 100 markers per morgan, if effective size (N_e) is 100. Marker characteristics and linkage disequilibria were obtained by simulating a population over 1,000 generations to achieve a mutation drift balance. The marker designs were evaluated for their accuracy of predicting breeding values from either estimating marker effects or estimating effects of haplotypes based upon combining 2 markers. Using microsatellites as direct marker effects, the accuracy of selection increased from 0.63 to 0.83 as the density increased from $0.25N_e$/morgan to $2N_e$/morgan. Using SNP markers as direct marker effects, the accuracy of selection increased from 0.69 to 0.86 as the density increased from $1N_e$/morgan to $8N_e$/morgan. The SNP markers required a 2 to 3 times greater density compared with using microsatellites to achieve a similar accuracy. The biases that genomic selection EBV often show are due to the prediction of marker effects instead of QTL effects, and hence, genomic selection EBV may need rescaling for practical use. Using haplotypes resulted in similar or reduced accuracies compared with using direct marker effects. In practical situations, this means that it is advantageous to use direct marker effects, because this avoids the estimation of marker phases with the associated errors. In general, the results showed that the accuracy remained responsive with small bias to increasing marker density at least up to $8N_e$ SNP/morgan, where the effective population size was 100 and with the genomic model assumed. For a 30-morgan genome and $N_e = 100$, this implies that about ~24,000 SNP are needed.

Key words: accuracy of selection, breeding value estimation, dense marker map, genome-wide selection, marker-assisted selection

J. Anim. Sci. 2008. 86:2447–2454
doi:10.2527/jas.2007-0010

INTRODUCTION

Information from high-density marker maps and high-throughput genotyping can be utilized in new selection methods. Current methodology for utilizing markers in breeding programs is marker-assisted selection (e.g., Meuwissen and Goddard, 1996). A limitation of using marker-assisted selection is the limited variance explained by the detected QTL due to the use of stringent significance tests in QTL detection.

Meuwissen et al. (2001) made a first step toward predicting a total genetic value using a genome-wide dense map of highly informative markers. The method was termed genomic selection, and the idea was to estimate the effects of all genes or chromosomal segments simultaneously. The effect of these segments is summed to predict the total breeding value. Selection can then be based on these breeding values. By treating the markers or haplotypes as random effects, the problem of estimating large numbers of haplotype effects from a limited number of animals can be managed. Meuwissen et al. (2001) compared different methods for predicting breeding values based on haplotype effects and found accuracies in the range of 0.79 to 0.85. Although the accuracy obtained depends on the genomic and phenotypic models assumed, these accuracies were obtained

[1]Corresponding author: trygve.roger.solberg@umb.no
Received January 4, 2007.
Accepted April 2, 2008.

only by phenotyping the parental and grandparental generation and were greater than the accuracy obtained if the breeding values of both parents had been known without error.

Factors affecting the accuracy of the prediction of the genotypes are largely unknown. At present it is unknown how dense markers need to be, particularly if they vary in information content. Therefore, the main aim of this paper was to examine how the accuracy of predicted breeding values responded to changes in marker densities with 2 different classes of markers such as microsatellites and SNP. Additionally, we examined if effects of marker genotypes should be used or if it was advantageous to use haplotype effects instead, due to their increased informativeness.

MATERIALS AND METHODS

Animal Care and Use Committee approval was not obtained for this study because no animals were used.

This study examined by simulation the changes in accuracy resulting from changing the type and density of markers and the approach to estimating effects with the evaluation. Throughout all comparisons, a Markov chain Monte Carlo approach was taken to the analysis with markers and QTL obtained from 1,000 generations of mutation selection balance.

Population

A stochastic model was simulated with a base generation of 100 unrelated animals (50 males and 50 females). To form generation 1 and later generations, sires and dams were mated randomly, but excluding selfing, for 1,000 generations. The effective population size ($\mathbf{N_e}$) was 100. The number of offspring born in generation t = 1,001 was increased to 1,000 using a factorial mating design involving all 50 sires and 50 dams from generation t = 1,000, where each sire was mated to 20 dams with 1 offspring per mating pair. In generation t = 1,002 a further 1,000 individuals were created by randomly sampling with replacement the sire and dam from among the 500 sires and 500 dams in generation t = 1,001.

Genome

In total, 10 chromosomes of equal length were simulated in the genome, such that the total genome size was 10 morgan. Markers were distributed evenly along the chromosomes but with differing densities. Marker densities are expressed throughout the paper in terms of N_e and morgan if not otherwise stated, because linkage disequilibrium is a function of $4N_ec$, where c = the distance between the loci, and thus the same linkage disequilibrium is obtained when doubling N_e and halving c (i.e., doubling marker density). For example, $8N_e$/morgan is equivalent to 800 markers per chromosome in our case, where N_e = 100, or 8,000 markers in total. In total, 4 different density schemes were evaluated for each of the 2 marker types. Using microsatellites, the 4 density schemes were $2N_e$/morgan, $1N_e$/morgan, $0.5N_e$/morgan, and $0.25N_e$/morgan, whereas, using SNP markers, the densities were $8N_e$/morgan, $4N_e$/morgan, $2N_e$/morgan, and $1N_e$/morgan. The total number of putative QTL, which can turn into a real QTL when a mutation occurs, was kept constant at 100 per chromosome for all marker density schemes. Marker and QTL positions are illustrated in Table 1 for 1 chromosome for the different marker density schemes. For example, with a density of $2N_e$/morgan, the chromosome begins with 2 markers, then 1 putative QTL, followed by 2 markers, a putative QTL, and so forth, in total 302 loci per chromosome.

Creation of Microsatellites and QTL

At t = 0 there was no polymorphism in marker or QTL loci. Each generation, for each individual at each locus, a random number was drawn in the simulation to test for the occurrence of a mutation. The mutation rate for the marker positions was 2.5×10^{-3} per locus per generation. The mutation rate for the QTL positions was 2.5×10^{-5} per locus per generation, and whether a putative QTL caused genetic variance or not depended on the mutations at these putative QTL positions. For each new mutation at a QTL, an allelic effect was drawn from the gamma distribution with a shape parameter ß = 0.4 and a scale parameter of

Table 1. Illustrated marker (M) and QTL (Q) positions[1] on 1 chromosome for different marker density schemes

Marker density[2]	Illustrated marker and QTL position on 1 chromosome
$0.25N_e$/morgan	M_1-Q_1-Q_2-Q_3-Q_4-M_2-…//…-M_{25}-Q_{97}-Q_{98}-Q_{99}-Q_{100}-M_{26}
$0.5N_e$/morgan	M_1-Q_1-Q_2-M_2-…//…-M_{50}-Q_{99}-Q_{100}-M_{51}
$1N_e$/morgan	M_1-Q1-M_2-…//…-M_{100}-Q_{100}-M_{101}
$2N_e$/morgan	M_1-M_2-Q_1-M_3-M_4-…//…-M_{199}-M_{200}-Q_{100}-M_{201}-M_{202}
$4N_e$/morgan	M_1-M_2-M_3-M_4-Q_1-M_5-M_6-M_7-M_8-…//…-M_{397}-M_{398}-M_{399}-M_{400}-Q_{100}-M_{401}-M_{402}-M_{403}-M_{404}
$8N_e$/morgan	M_1-M_2-M_3-M_4-M_5-M_6-M_7-M_8-Q_1-M_9-M_{10}-M_{11}-M_{12}-M_{13}-M_{14}-M_{15}-M_{16}-…//…-M_{793}-M_{794}-M_{795}-M_{796}-M_{797}-M_{798}-M_{799}-M_{800}-Q_{100}-M_{801}-M_{802}-M_{803}-M_{804}-M_{805}-M_{806}-M_{807}-M_{808}

[1]M_x represents a marker, whereas the Q_x symbolizes a putative QTL. One chromosome is 100 cM in length, and in total, 10 chromosomes were simulated, such that the genome size was 10 morgan.

[2]Marker density indicates the number of markers in terms of effective population size (N_e) and the distance (in morgan).

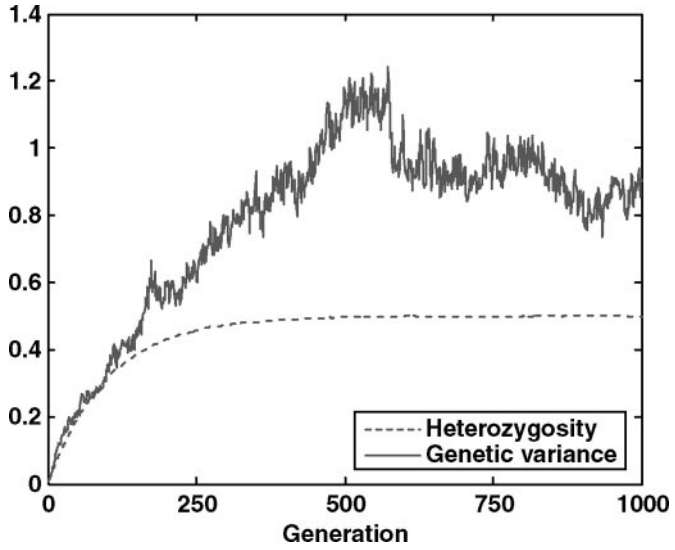

Figure 1. Genetic variance and marker heterozygosity using a density of $1N_e$/morgan between microsatellite markers after 1,000 generations (N_e = 100). Lines are average genetic variance and heterozygosity for microsatellite markers based on 20 replicated simulations. N_e = effective population size.

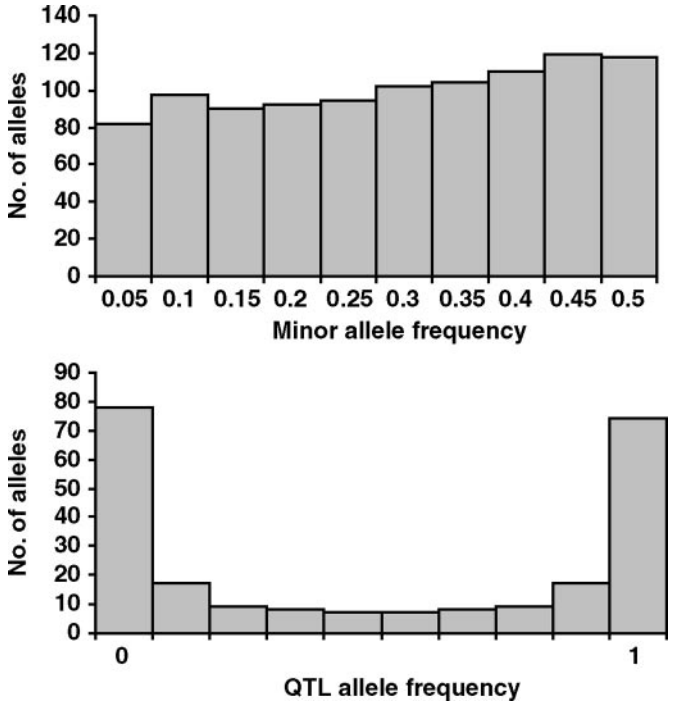

Figure 2. Typical minor allele frequency distribution of the SNP alleles (top) and a typical QTL allele frequency distribution of the segregated QTL.

1.66 (Hayes and Goddard, 2001). The shape parameter was chosen to reflect the empirical distribution of the QTL effects, and the scale parameter was chosen such that the expected total genetic variance equaled 1. The gamma distribution yields only positive effects; therefore, the QTL effect was sampled to be positive or negative with probability 0.5. Typically, the number of segregating QTL was between 5 and 6% of the total number of QTL, and the distribution of the QTL allele frequencies resembled a U-shaped distribution as in the Wright-Fisher mutation drift equilibrium (Wright, 1931, 1935). Figure 1 shows the total genetic variance and heterozygosity of microsatellite markers plotted against the number of generations.

SNP Markers

The SNP markers were obtained by recoding microsatellite alleles. First, the evolutionary tree of how the microsatellite alleles evolved was stored (e.g., allele 4 may have come from a mutation in allele 2 and so on). Second, one of the mutations in this evolutionary tree was assumed visible and the others were invisible, which resulted in only 2 alleles, 1 mutated allele and 1 ancestral allele. The mutation to be visible was chosen such that the SNP allele frequency was as close as possible to 0.5 in generations t = 1,001 and 1,002. Figure 2 shows a typical minor allele frequency of the SNP markers, illustrated using the $1N_e$/morgan density scheme. The minor allele frequencies of the SNP markers showed approximately a uniform distribution with an overrepresentation of marker alleles with intermediate frequencies, which reflects the effect in practice of prescreening SNP markers and selecting the most informative. This resulted in a typical number of segregating SNP markers of 98 to 99% of the total number of markers.

Phenotypes

Only the 1,000 individuals in generation t = 1,001 were assumed to have a phenotypic record. Phenotypic records in generation t = 1,001 were obtained by $\mathbf{P}_i = \mathbf{TBV}_i + \boldsymbol{\varepsilon}_i$, where $\mathbf{TBV}_i$ = the true breeding value of the ith animal and $\boldsymbol{\varepsilon}_i$ sampled from $N(0, \sigma_e^2)$. The environmental variance (σ_e^2) was set equal to the true genetic variance (σ_g^2); therefore, the heritability was 0.5 for every replicate. The genetic variance varied somewhat from replicate to replicate, but was on average close to 1 (Figure 1).

The effect of doubling the number of phenotypes was tested by doubling the number of animals in t = 1,001 to 2,000 using the factorial mating design described above, except that each sire was mated to 40 dams with 1 offspring per mating pair. In this scenario, a further 2,000 unphenotyped individuals were created in t = 1,002 by randomly sampling with replacement its sire and dam among the 1,000 sires and 1,000 dams in generation t = 1,001.

Estimation of Marker and Haplotype Effects

The markers were treated either as marker genotype effects or as 2 markers combined into a haplotype. When haplotypes are referred to in this text, it means

2 neighboring markers combined into a haplotype, unless otherwise stated. Haplotypes used densities of $1N_e$/morgan and $2N_e$/morgan using microsatellites and $4N_e$/morgan and $8N_e$/morgan using SNP markers. Recombination between the markers followed Haldane's mapping function (Haldane, 1919). In each simulated population, the BayesB method of Meuwissen et al. (2001) was used to estimate the effects of single markers and haplotypes in generation t = 1,001. The model at the level of the data was: $\mathbf{y} = \mu\mathbf{1}_n + \Sigma_i\mathbf{X}_i\mathbf{g}_i + \mathbf{e}$, where $\mathbf{y}$ = the vector of phenotypes; $\mathbf{1}_n$ = a vector of n ones; Σ_i = the summation over all markers (haplotypes); $\mathbf{X}_i$ = a design matrix for the ith marker; $\mathbf{g}_i$ = the vector of marker (haplotype) effects; and $\mathbf{e}$ = the error. The variance of the marker effects is $\sigma^2_{gi,}$ which is estimated for every marker using an informative prior distribution. The distribution of the genetic variances across loci resembles a situation where there are many loci with no genetic variance (not segregating) and some with genetic variance. The prior distribution, therefore, is a mixture of distributions with a probability π, at $\sigma^2_{gi} = 0$ and an inverted chi-square distribution $\chi^{-2}(v, S)$ with probability (1-π) for $\sigma^2_{gi} > 0$. The parameters of the prior distribution were v = 4.234 and S = 0.0429 (Meuwissen et al., 2001). The probability π depends on the density of the markers and varies with different marker densities, because with more markers, it becomes less likely for marker i to be closest to a QTL. At 0.25, 0.5, 1, 2, 4, and $8N_e$/morgan, the values of π were 0.212, 0.106, 0.053, 0.0265, 0.01325, and 0.006625, respectively.

Sampling from a prior for σ^2_{gi}, which was a mixture distribution, was by a Metropolis-Hastings algorithm that sampled σ^2_{gi} from $p(\sigma^2_{gi} | \mathbf{y}^*)$, where the prior distribution, $p(\sigma^2_{gi})$ was used as the distribution to suggest updates for the Metropolis-Hastings chain (Gilks et al., 1996) and $\mathbf{y}^*$ denotes the data $\mathbf{y}$ corrected for the mean and all other genetic effects except the marker (haplotype) effect ($\mathbf{g}_i$). The Metropolis Hastings chain was run for 10,000 cycles using a burn-in period of 1,000 cycles. Given σ^2_{gi}, marker (haplotype) effects, $\mathbf{g}_i$, were sampled from $p(\mathbf{g}_i | \sigma^2_{gi})$ using Gibbs sampler (Sørensen and Gianola, 2002).

Linkage Disequilibrium and Marker Informativeness

Linkage disequilibrium (**LD**) was estimated as the average r^2 value for 2 adjacent SNP markers using 10 replicates for all 1,000 animals in generation t = 1,001 for the 4 different SNP density schemes. The formulae used to calculate LD was $r^2 = D^2/(p_1p_2q_1q_2)$, where D = the coefficient of disequilibrium. The frequency of allele 1 and 2 in locus 1 and frequency of allele 1 and 2 in locus 2, are p_1, $p_2 = (1-p_1)$, q_1 and $q_2 = (1-q_1)$, respectively. Figure 3 shows the estimated r^2 value for the 4 different density schemes when adjacent SNP markers were evaluated. The estimated LD was similar to the expected value of LD if a population is in recombination drift balance and allows for mutations, which is approximately $1/(2 + 4N_ec)$, where c = the recombination rate between the markers (Tenesa et al., 2007). The LD increased with increasing marker density, as expected. Marker informativeness was calculated for both microsatellites and SNP markers using the polymorphism information content (**PIC**; Lynch and Walsh, 1998).

Prediction and Accuracy of Breeding Values

Generation t = 1,002 was not phenotyped, but breeding values in generation t = 1,002 were estimated using the phenotypic records of t = 1,001 and the genomic records in t = 1,001 and 1,002. The EBV were made either using marker genotypes or combining neighboring marker genotypes into haplotypes. The EBV were compared with the TBV in generation t = 1,002. Estimated breeding values of animal j were obtained after estimating the marker-haplotype effects from:

$$\mathrm{EBV}_j = \sum_i^n \mathbf{X}_{ji}\mathbf{g}_i,$$

where $\mathbf{X}_{ji}$ denotes the marker genotype of animal j at locus i in generation t = 1,002 and $\mathbf{g}_i$ = the estimate of the marker or haplotype effects, which were estimated on animals in generation t = 1,001.

Twenty replicated simulations were performed per marker density and marker type as described above. A regression analysis was performed of TBV on estimated breeding values. The regression analysis resulted in a regression coefficient (b) and a correlation coefficient (r), where the regression coefficient reflects the bias of the breeding value estimate, b = 1 denotes unbiased estimates, and the correlation coefficient reflects the accuracy of predicting the breeding values.

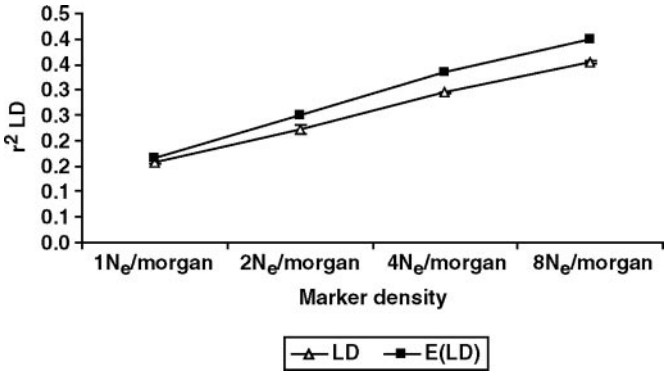

Figure 3. Calculated r^2 linkage disequilibrium (LD) and expected r^2 LD [E(LD)] in generation t = 1,001 for adjacent SNP markers for the 4 different marker density schemes. N_e = effective population size.

Table 2. Typical marker informativeness[1] for microsatellite and SNP markers, illustrated with the polymorphism information content (PIC)

Marker type	PIC_{mean} ± SE	Average number of alleles per locus
Microsatellite	0.459 ± 0.0140	6.1
SNP	0.234 ± 0.0008	2.0

[1]The measured values are from generation t = 1,001 using a density of $1N_e$/morgan. N_e = effective population size.

RESULTS

Marker Informativeness

Typically, the number of microsatellite alleles was between 1 and 14, with an average of 6 alleles per locus genome-wide. Table 2 shows the marker informativeness (PIC) for microsatellites and SNP in generation t = 1,001. Mean values for the PIC genome-wide for microsatellites and SNP markers were typically 0.459 and 0.234, respectively. These can be compared with theoretical maxima for PIC of 0.810 for a 6-allele microsatellite and 0.375 for a biallelic SNP, where the maxima are achieved with equal frequency among the alleles at a locus.

Direct Estimation of Marker Effects

Four schemes with different densities for microsatellites and SNP markers were compared. The regression coefficients (b) of TBV on EBV and the accuracy of selection (i.e., the correlation between TBV and EBV) are given in Tables 3, 4, and 5. The accuracy of the EBV varied from 0.626 to 0.827 for microsatellites using a marker density of $0.25N_e$/morgan and $2N_e$/morgan, respectively (Table 3). Hence, the accuracy of estimating the breeding values using microsatellites increased as the density of the markers increased, as expected. The regression of TBV on EBV become closer to 1 as the marker density increased, although for SNP, this relationship was less clear.

From Tables 3 and 4, it is seen that the accuracy using $1N_e$/morgan for SNP markers is intermediate between the accuracies achieved with densities of $0.25N_e$/morgan and $0.5N_e$/morgan for microsatellite markers, indicating the need for a 2 to 3 times greater density when using SNP markers. However, for greater densities of microsatellites, this equivalence is closer to a 2-fold density of SNP.

Table 3. Accuracy of selection (r) and regression coefficient (b) of true breeding value (TBV) on EBV when EBV are estimated using microsatellite genotypes

Marker density[1]	$r_{TBV;EBV}$ ± SE	$b_{TBV;EBV}$ ± SE
$0.25N_e$/morgan	0.626 ± 0.014	0.835 ± 0.015
$0.5N_e$/morgan	0.723 ± 0.010	0.882 ± 0.013
$1N_e$/morgan	0.770 ± 0.013	0.882 ± 0.013
$2N_e$/morgan	0.827 ± 0.010	0.941 ± 0.012

[1]N_e = effective population size.

Table 4. Accuracy of selection (r) and regression coefficient (b) of true breeding value (TBV) on EBV when EBV are estimated using SNP genotypes

Marker density[1]	$r_{TBV;EBV}$ ± SE	$b_{TBV;EBV}$ ± SE
$1N_e$/morgan	0.690 ± 0.016	0.877 ± 0.019
$2N_e$/morgan	0.790 ± 0.07	0.879 ± 0.011
$4N_e$/morgan	0.841 ± 0.004	0.943 ± 0.009
$8N_e$/morgan	0.860 ± 0.010	0.923 ± 0.011

[1]N_e = effective population size.

In general, the accuracies for both microsatellites and SNP markers increased about 1.04 to 1.07-fold when the marker density was doubled. Figure 4 summarizes the results graphically, where the average accuracy for the 20 replicates is plotted against the marker density when using marker genotypes for estimation.

When the number of offspring was doubled to 2,000 animals, the accuracy increased 1.09-fold from 0.77 to 0.84 using microsatellites and a density of $1N_e$/morgan. The EBV were also less biased when the number of animals increased, as the regression coefficient increased from 0.88 to 0.93 (Table 5).

Using Haplotypes to Estimate Breeding Values

Two different densities using haplotypes for both microsatellites and SNP were evaluated. The regression coefficients (b) of TBV on EBV and the accuracy of EBV are given in Tables 6 and 7.

When haplotypes were obtained from microsatellite markers, the accuracy varied from 0.764 to 0.798 using densities of $1N_e$/morgan and $2N_e$/morgan, respectively (Table 6). Microsatellite haplotypes gave similar ac-

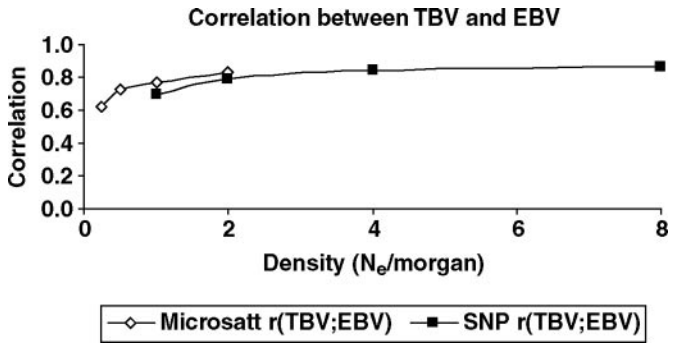

Figure 4. Average correlation coefficient (20 replicates; i.e., accuracy of selection) for the 4 different density schemes using marker genotype effect for microsatellite (Microsatt) and SNP markers. TBV = true breeding value. N_e = effective population size.

Table 5. Accuracy of selection (r) and regression coefficient (b) of true breeding value (TBV) on EBV using microsatellite markers with a density of $1N_e$/morgan, when the number of offspring was doubled[1]

Coefficient	1,000 offspring	2,000 offspring
$r_{TBV;EBV}$ ± SE	0.770 ± 0.013	0.842 ± 0.010
$b_{TBV;EBV}$ ± SE	0.882 ± 0.013	0.930 ± 0.010

[1]N_e = effective population size.

curacies, at least not significantly different compared with using microsatellite genotype effects, although the tendency was to a reduction in accuracy using haplotypes. The bias was greater when using haplotypes. These trends were repeated when comparing SNP haplotypes and SNP genotypes.

DISCUSSION

This study demonstrates the effect of marker density and type on the prediction of breeding values for the next generation, when the single marker effects or haplotype effects have been estimated in a previous generation. When using SNP markers for genomic selection, a 2 to 3 times greater density was required compared with using microsatellites to achieve comparable accuracy using the same genetic architecture. A density of $2N_e$/morgan using microsatellites or $4N_e$/morgan using SNP yielded an accuracy of selection >0.8 for populations with an effective size of 100, 1,000 phenotypes, and a heritability of 0.5. Using haplotypes resulted in similar or reduced accuracies compared with using marker genotypes. In practice, it is therefore advantageous to use marker genotypes, because this avoids the estimation of marker phases and errors associated with this.

It is perhaps counterintuitive that the use of marker haplotypes instead of marker genotypes yielded somewhat lower accuracies at a relatively low SNP density of $4N_e$/morgan, and similar tendencies were found for the microsatellite markers (Tables 6 and 7). The covariance structure imposed by fitting marker genotypes differs from that by fitting marker haplotypes. For example, for 2 adjacent markers M and N with M11N11 denoting the marker genotypes at M and N: when fitting marker genotype, there is a covariance between records y_{M11N11} and y_{M11N22}, due to the common genotype at M, whereas there is no covariance between these records when marker haplotypes are fitted. Apparently, assuming covariance between records that carry the same marker genotype yields better prediction of the QTL genotypes than assuming this covariance is absent. If we assume that one of the markers is the better predictor of the QTL, say marker M, it is apparently not always beneficial to make haplotypes with an adjacent marker, which itself may not be a good predictor of the QTL. The latter is especially the case when the adjacent markers are further apart (i.e., having relatively low r^2), and thus their r^2 with the QTL may be different (Table 7).

Table 6. Accuracy of selection (r) and regression coefficient (b) of true breeding value (TBV) on EBV when EBV are estimated using microsatellite haplotypes

Marker density[1]	$r_{TBV;EBV}$ ± SE	$b_{TBV;EBV}$ ± SE
$1N_e$/morgan	0.764 ± 0.010	0.847 ± 0.015
$2N_e$/morgan	0.798 ± 0.011	0.839 ± 0.014

[1]N_e = effective population size.

Table 7. Accuracy of selection (r) and regression coefficient (b) of true breeding value (TBV) on EBV when EBV are estimated using SNP haplotypes[1]

Marker density	$r_{TBV;EBV}$ ± SE	$b_{TBV;EBV}$ ± SE
$4N_e$/morgan	0.802 ± 0.011	0.891 ± 0.012
$8N_e$/morgan	0.821 ± 0.015	0.902 ± 0.012

[1]N_e = effective population size.

A simulation study, as carried out here, requires several assumptions about the underlying genetic model and the population history. The genetic model assumed here is a model, just like the infinitesimal model is a model, and the assumptions will never hold exactly in any practical situation. However, we believe that our main conclusions about the effects of marker density and types of markers are general and they will qualitatively also hold for different genetic models. Nevertheless, it may be that different genetic models (e.g., with fewer large QTL) may require adjusting the prior distribution of the genetic model at hand. It is important that the prior distribution and the true underlying genetic model correspond well, and this will require research on the distribution of the QTL effects (e.g., Hayes and Goddard, 2001). The prior distribution of the variances of marker effects used here was a mixture distribution of an inverted chi-square and a distribution with zero variance, whereas the true QTL effects were sampled from the gamma distribution. Hence, the prior distribution used for the analysis and the distribution for the sampling of the true QTL effects did not agree exactly in this study, which may explain the bias observed in the EBV (b values <1 in Tables 3 and 4). However, we also simulated 5 replicates, where it was assumed that the QTL genotypes at the true QTL positions were known (i.e., running the model using only QTL genotypes) and the estimates of the QTL effects were used to predict the EBV. The accuracy of selection in that case was 0.919 ± 0.013, and the regression of TBV on EBV was 1.000 ± 0.023, which indicates that in our study the bias came from the use of marker effects rather than from the distributional assumptions made concerning QTL. The accuracy of 0.919 also places an upper bound for the accuracy that can be expected from using a very dense marker map. In practice, we have to be aware of possible biases of genomic selection EBV,

Table 8. Comparison of accuracy of selection (r) and regression (b) of true breeding value (TBV) on EBV when EBV are estimated using SNP markers when N_e = 200 using a marker density of $1N_e$/morgan compared with using N_e = 100 and a marker density of $1N_e$/morgan[1]

Marker density	$r_{TBV;EBV}$	$b_{TBV;EBV}$
N_e = 100, $1N_e$/morgan	0.690 ± 0.016	0.877 ± 0.019
N_e = 200, $1N_e$/morgan	0.677 ± 0.012	0.890 ± 0.026

[1]N_e = effective population size.

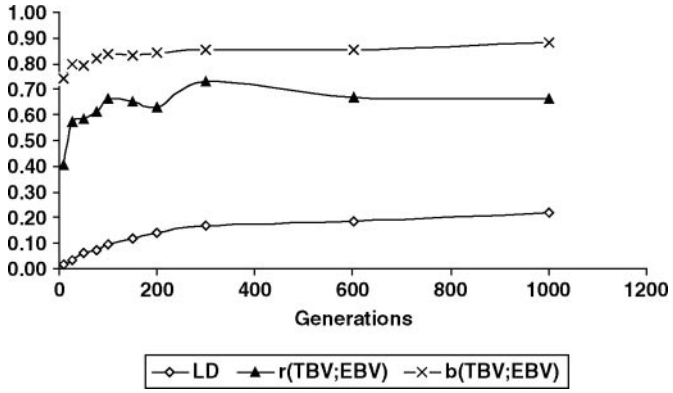

Figure 5. Accuracy of selection (r) and regression of true breeding value (TBV) on EBV (b) when the amount of linkage disequilibrium (LD) gradually increased in the population. Based on 6 replicates using SNP markers with a density of $1N_e$/morgan. N_e = effective population size.

which may be due to differences between the prior distribution used in the EBV estimation and the true underlying genetic model and from using markers instead of QTL effects. These biases may be corrected for using a cross-validation type of approach, in which some phenotypes are hidden from the analysis and later the regression of the hidden phenotypes on the predicted EBV is estimated. This regression coefficient should equal 1, and if this is not the case, the EBV can be rescaled such that the regression coefficient equals 1.

Although there are many structural similarities in the simulated populations and the parameters used in this study compared with Meuwissen et al. (2001), the results show that there is no barrier to achieving such accuracies using genomic evaluations in practice. In a traditional progeny test scheme in Canadian Holsteins, the accuracy of predicting the EBV of progeny-tested young bulls for production, conformation, fertility, and longevity was estimated to be 0.75 for their first EBV (Schaeffer, 2006). Compared with this, a density of $1N_e$/morgan using microsatellites or $2N_e$/morgan using SNP seems sufficient to achieve a similar accuracy.

The presented simulations assumed a relatively small effective population size of N_e = 100, which generates LD between the markers and a QTL and thus causes the marker effects. The expected amount of disequilibrium in a stable population represents a balance between its creation by drift and its decay by recombination. For a randomly mating population in which the drift-recombination balance has been achieved, the expected squared correlation between the presence of 2 linked loci is $E(r^2) = 1/(1 + 4N_ec)$, where c = the recombination between 2 loci (Lynch and Walsh, 1998). Thus, if N_e is 2 times greater than in our simulations, the marker density needs to be doubled to generate the same LD and thus to have the same accuracy of selection. To test this, a simulated population with N_e = 200 using SNP markers with a marker density of $1N_e$/morgan (i.e., N_e = 200) was compared with a population using N_e = 100 and a marker density of $1N_e$/morgan. Table 8 shows that using N_e = 200 gave almost the same accuracy compared with using N_e = 100 using half the marker density. Note that in both schemes, marker densities are $1N_e$/morgan, demonstrating that selection accuracies are similar for similar densities expressed in N_e/morgan units, which is why we used these units throughout this paper.

The LD is the key factor that is driving the genomic prediction process, and, to further confirm this, some simulations were tested to see how the accuracy changed with increasing LD. The simulated population was therefore stopped before it reached a balance between recombination and drift, and before the equilibrium amount of LD was reached, resulting in a lower LD. Figure 5 shows that the increase in LD and the increase of the accuracy curve are very similar, suggesting that LD is indeed driving the accuracy of selection.

A heritability of 0.5 was used in this simulation study. However, many traits in animal breeding show a smaller heritability than 0.5 (i.e., they will show relatively more environmental variance). A reduced heritability will lead to a decrease in accuracy of predicting the breeding value but can be compensated for by using a larger number of observations to estimate the marker (haplotype) effects.

The estimation model assumes that there is no dominance (i.e., only the additive effects are fitted), and the average effects of the genes are estimated, which is probably satisfactory for the prediction of breeding values in most cases. When the prediction of dominance effects is important to predict total genetic values, dominance effects need to be added to the statistical model. In theory, this is not a problem, but research is needed to verify that such estimates are accurate in realistic scenarios. Another assumption in the simulation model is that the markers and QTL are evenly distributed on the chromosome, which in fact is realistic for microsatellite markers (Liu and Cordes, 2004).

This study showed that the results of Meuwissen et al. (2001) could be extended to SNP markers, which make dense high-throughput genotypes possible. At greater densities, one needs about twice as many SNP as microsatellites. Using such dense SNP genotyping technology may make selection for complex traits, or traits that are not widely recorded, possible, by esti-

mating the marker effects in one generation, and using these effects in later generations to select their descendents.

LITERATURE CITED

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Marcov Chain Monte Carlo in Practice. Chapman & Hall/CRC, Boca Raton, FL.

Haldane, J. B. S. 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. J. Genet. 8:299–309.

Hayes, B. J., and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33:209–229.

Liu, Z. J., and J. F. Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture 238:1–37.

Lynch, M., and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates Inc, 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.

Meuwissen, T. H. E., B. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223.

Sørensen, D., and D. Gianola. 2002. Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. Springer-Verlag, New York, NY.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17:520–526.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Wright, S. 1935. Evolution in populations in approximate equilibrium. J. Genet. 30:257–266.