Genomic selection in admixed and crossbred populations[1]

A. Toosi, R. L. Fernando,[2] and J. C. M. Dekkers

Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames 50011

ABSTRACT: In livestock, genomic selection (GS) has primarily been investigated by simulation of purebred populations. Traits of interest are, however, often measured in crossbred or mixed populations with uncertain breed composition. If such data are used as the training data for GS without accounting for breed composition, estimates of marker effects may be biased due to population stratification and admixture. To investigate this, a genome of 100 cM was simulated with varying marker densities (5 to 40 segregating markers per cM). After 1,000 generations of random mating in a population of effective size 500, 4 lines with effective size 100 were isolated and mated for another 50 generations to create 4 pure breeds. These breeds were used to generate combined, F_1, F_2, 3- and 4-way crosses, and admixed training data sets of 1,000 individuals with phenotypes for an additive trait controlled by 100 segregating QTL and heritability of 0.30. The validation data set was a sample of 1,000 genotyped individuals from one pure breed. Method Bayes-B was used to simultaneously estimate the effects of all markers for breeding value estimation. With 5 (40) markers per cM, the correlation of true with estimated breeding value of selection candidates (accuracy) was greatest, 0.79 (0.85), when data from the same pure breed were used for training. When the training data set consisted of crossbreds, the accuracy ranged from 0.66 (0.79) to 0.74 (0.83) for the 2 marker densities, respectively. The admixed training data set resulted in nearly the same accuracies as when training was in the breed to which selection candidates belonged. However, accuracy was greatly reduced when genes from the target pure breed were not included in the admixed or crossbred population. This implies that, with high-density markers, admixed and crossbred populations can be used to develop GS prediction equations for all pure breeds that contributed to the population, without a substantial loss of accuracy compared with training on purebred data, even if breed origin has not been explicitly taken into account. In addition, using GS based on high-density marker data, purebreds can be accurately selected for crossbred performance without the need for pedigree or breed information. Results also showed that haplotype segments with strong linkage disequilibrium are shorter in crossbred and admixed populations than in purebreds, providing opportunities for QTL fine mapping.

Key words: admixture, crossbreeding, genomic selection, marker-assisted selection

J. Anim. Sci. 2010. 88:32–46
doi:10.2527/jas.2009-1975

INTRODUCTION

Genomic selection (**GS**; Meuwissen et al., 2001) is a form of marker-assisted selection that uses marker genotypes and phenotypes in a training population to simultaneously estimate effects of a large number of markers across the genome for the purpose of predicting breeding values (**BV**) of selection candidates based on their marker genotypes. The accuracy of GS depends on the amount of linkage disequilibrium (**LD**) between QTL and markers and the number of records available to estimate marker effects. Most commercial beef cattle populations consist of animals with different and often unknown breed compositions. The presence of an unknown population structure has raised concerns about using admixed or crossbred populations as training data for GS, yet these are the populations that are most relevant as the target for genetic improvement of purebreds (Dekkers, 2007).

Admixture is the presence of multiple genetically distinct subgroups within a population (Wang et al., 2005). Numerous studies (e.g., Rabinowitz, 1997; Flint-Garcia et al., 2003; Hirschhorn and Daly, 2005) have reported that admixture can produce spurious associa-

[1]This research was motivated by a question by R. L. Quaas (Cornell University, Cornell, NY) at a meeting of the statistical methods group of the National Beef Cattle Evaluation Consortium. Funding for this research was provided by Newsham Choice Genetics, the United States Department of Agriculture, National Research Initiative grant USDA-NRI-2007-35205-17862, and the Iowa Agricultural and Home Economics Experiment Station, Ames.

[2]Corresponding author: rohan@iastate.edu

Received March 20, 2009.
Accepted August 31, 2009.

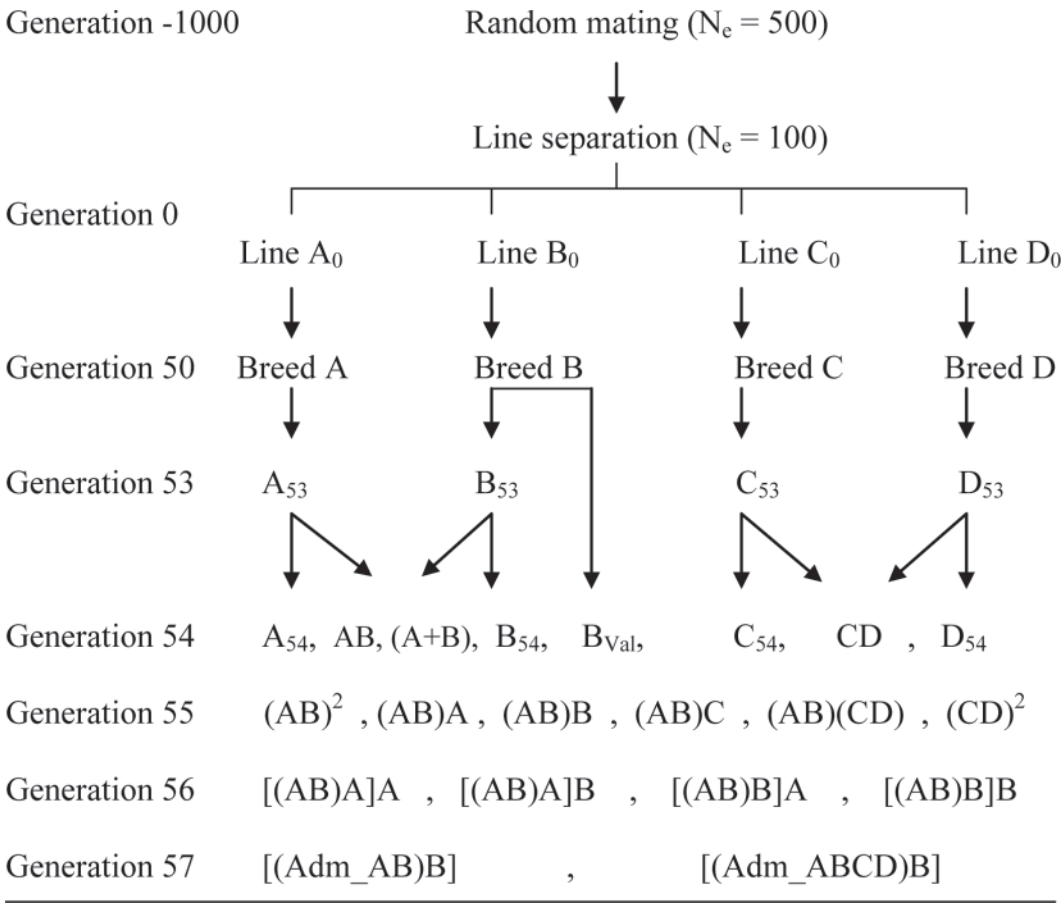

Figure 1. Schematic representation of the simulated population history (N_e = effective population size) and the different types of crossbred and admixed populations that were simulated. A_{54} and B_{54} represent purebred training populations; AB is an F_1 training population; A+B is a training population consisting of individuals from breeds A and B; $(AB)^2$ is an F_2 training population; (AB)A, (AB)B, [(AB)A]A, [(AB)A]B, [(AB)B]A, and [(AB)B]B are different back-cross populations; [(Adm_AB)B] and [(Adm_ABCD)B] are admixed training populations of 2 and 4 breeds; B_{Val} is the validation population. As of generation 54, arrows are not shown for simplicity of the picture.

tions and seriously elevate false discovery rates in QTL detection. Several methods have been proposed to address this problem (e.g., Kennedy et al., 1992; Spielman et al., 1993; Pritchard et al., 2000; Meuwissen et al., 2002; Price et al., 2006; Yu et al., 2006).

Ideally, however, if all QTL that explain genetic variation in the trait of interest were included in the model, it would not be necessary to explicitly account for pedigree or breed composition in the analysis. Thus, provided high-density SNP are used and analyzed simultaneously, as in GS, pedigree and breed composition need not be explicitly modeled. The objective of this work, therefore, was to evaluate accuracy of GS with high-density markers for predicting BV of purebred animals based on estimates of marker effects in a crossbred or admixed population, without explicitly accounting for pedigree or breed composition.

MATERIALS AND METHODS

No live animals were used for this study, and therefore, institutional animal care and use committee approval was not required.

Population

A base population of unrelated individuals was stochastically simulated and used as the ancestral population of 4 pure breeds that were used to create admixed and crossbred populations (Figure 1). The base population was randomly mated for 1,000 generations, including selfing, with an effective size ($\mathbf{N_e}$) of 500. To simulate the 4 purebred populations (referred to as breeds A, B, C, and D hereafter), at generation zero, 4 independent random samples of 100 animals were drawn from the base population, and each was randomly mated (including selfing) for another 53 generations, with N_e of 100.

Breeds A and B were then crossed to produce an F_1 (AB) population. Mating (AB) with breeds A and B created backcrosses (AB)A and (AB)B, which were again backcrossed to A or B to create crossbred populations [(AB)A]A, [(AB)A]B, [(AB)B]A, and [(AB)B]B. An F_2 population, $(AB)^2$, was created by inter-mating the (AB) F_1. Similarly, breeds C and D were used to create corresponding crossbred populations.

In beef cattle, commercial animals are often produced by mating purebred sires to crossbred dams of hetero-

Table 1. The parameters used for the simulation program

Item	
Genome size	100 cM
Number of chromosomes	1
Marker density per cM	5, 10, 20, or 40
Number of segregating QTL	100
Mutation rate of QTL or marker locus	2.5×10^{-5}
Minor allele frequency	≥0.10
Distribution of additive QTL effects	Gamma (shape = 0.4; scale = 1/1.66)
Prior distribution for variance of nonzero marker effects (σ^2_{gi})	χ^{-2} (ν = 4.234, S = 0.0429)
π[1]	0.950 for marker density ≤20 and 0.975 for marker density of 40 per cM
Population size	
Generations −1,000 to 0	N_e[2] = 500
Generations 0 to 53	N_e = 100
Generation ≥54	N[3] = 1,000
Heritability	0.30
Residual variance	1.00

[1]π is probability ($\sigma^2_{gi} = 0$) for Bayes-B method.
[2]Effective population size.
[3]Number of phenotypic or genotypic records, or both.

geneous breed composition. To simulate such a population, an admixed population of 2 breeds [(Adm_AB) B] was created by first putting dams from breed A and all crossbreds involving breeds A and B in the same group and then mating them to sires from breed B. An admixed population of 4 breeds [(Adm_ABCD)B] was formed by mating dams from all breeds, excluding breed B, and all of the crosses involving the 4 breeds to sires from breed B. Further, 3-way and 4-way crosses were made by crossing (AB) with C and (AB) with (CD). To create the combined population, 2 random samples of equal size from purebreds A and B were put together into a single population. In the remainder, this latter population will be referred to as the combined_AB population, in contrast to the admixed populations described previously, which include purebreds and their crosses.

Each of the purebreds A and B, the (AB), $(AB)^2$, (AC), (AB)C, (AB)(CD) crossbreds, combined_AB, and admixed data sets was used as a training population consisting of 1,000 animals for estimating marker effects. These animals were created by mating randomly sampled individuals from the appropriate parental lines, each of size 100. Thus, on the average, each sire (and dam) had 10 offspring in the training data set. Because the objective was to determine how well marker effects estimated in the various training populations predicted BV of purebred individuals, a separate generation of purebred population B was used for validation. To generate the validation population of size 1,000, in generation 50 a sample of 1,000 animals was drawn from breed B and randomly mated for another 4 generations (B_{Val} in Figure 1).

To evaluate the impact of breed differences on the accuracy of GS, a second scenario was also considered, in which breeds were separated in generation 25 rather than generation 0 (Figure 1), such that breeds were diverged for only 25 instead of 50 generations. To maintain the same level of LD, effective population size was reduced from 500 to 100 in generation 0, as before.

Genome

To make the simulation computationally feasible, a genome consisting of 1 chromosome of 100 cM with 100 segregating QTL and different marker densities was simulated (Table 1). To end up with the required number of segregating loci after 1,000 generations, about 3 times as many biallelic loci were simulated with starting allele frequencies of 0.5 and a reversible random mutation rate of 2.5×10^{-5}. Each locus was a marker locus or a QTL. A binomial map function was used to simulate recombination, and interference was allowed by setting the maximum number of uniformly and independently distributed crossovers on the chromosome to be 4 (Karlin, 1984). To make a marker panel, 500, 1,000, 2,000, or 4,000 marker loci were drawn at random from segregating loci, minor allele frequency (**MAF**) ≥ 0.10, at generation 53 after pooling all 4 breeds into a single cohort. As a result, markers were not evenly dispersed along the chromosome, and some of them may not be segregating in all training populations.

Phenotypes

To create phenotypic values for each training population, 100 QTL were randomly picked from the set of segregating QTL in that population. This was done in generation 54 for breed B, (AB), and combined_AB, in generation 56 for $(AB)^2$, 3- and 4-way cross data sets, and in generation 57 for the admixed training data sets (Figure 1). Note that QTL with MAF <0.10 in each breed will have an intermediate allele frequency in the combined_AB, crossbreds, or admixed training populations. The QTL were additive, and their effects were sampled from a gamma distribution with shape and

scale parameters of 0.4 and 1/1.66, respectively. This provides us with an L-shaped distribution of QTL effects, which Hayes and Goddard (2001) and Xu (2003) suggest is close to the real distribution of the QTL effects. With equal probability, 1 of the 2 alleles was chosen to be positive or negative. To keep the genetic variance constant across training populations, the effect of each QTL was scaled in each replicate. This was done to ensure that each training data set had the same genetic variance, such that this could not contribute to differences between training data sets.

The scaled QTL effects then were summed over all QTL genotypes for each individual to compute its true BV. With this setting, each training population received a different set of QTL affecting phenotypes, although the number of segregating QTL and the genetic variance were the same for all training populations. Finally, a standard normal deviate was added to each true BV to provide the phenotype of an individual for a quantitative trait with heritability 0.30.

It should be mentioned that here the whole genetic variance is assigned to a single chromosome, whereas in reality the total genetic variance is distributed to all chromosomes (30 chromosomes in the case of cattle, for instance). In this study, a short genome was chosen to reduce computational costs. An additional data set with a total of 5,000 markers and 100 QTL on 5 chromosomes, each of length 1 morgan, was simulated to examine the impact of genome size on our results. The analysis was run for the training populations of purebred B, [(Adm_ABCD)B], combined_AB and 4-way crossbred only, with 96 replicates.

Estimation of Marker Effects

Method Bayes-B of Meuwissen et al. (2001) was used to estimate effects of markers in the training data using the model: $\mathbf{y} = \mathbf{1}\mu + \Sigma_i \mathbf{x}_i g_i + \mathbf{e}$, where $\mathbf{y}$ is the vector of phenotypic values of individuals in the training data, μ is a single unknown population mean, $\mathbf{1}$ is a vector of ones, $\mathbf{x}_i$ is a column vector containing the genotypes (0, 1, or 2) of each individual at locus i, g_i is the random unknown allele substitution effect for marker i, and $\mathbf{e}$ is a random vector of unknown residuals with $e_i \sim N(0,\sigma^2_e)$. In method Bayes-B conditional on σ^2_{gi}, the gi has a point mass at 0, when $\sigma^2_{gi} = 0$ and has a $N(0,\sigma^2_{gi})$ distribution when $\sigma^2_{gi} > 0$. Further, the prior for σ^2_{gi} is a mixture distribution with a point mass at 0 with probability π and an inverted chi-squared distribution with known parameters $\nu = 4.234$ and $S = 0.0429$, with probability $(1 - \pi)$.

The probability π is assumed known. Based on preliminary analyses with several different π values, π was set equal to 0.95, except for a density of 40 markers per cM, for which π was increased to 0.975. A Markov chain Monte-Carlo (**MCMC**) of length 10,000 cycles with a burn-in period of 1,000 cycles was conducted. Convergence of the MCMC chain was examined using the R package CODA (Plummer et al., 2006).

Validation of Genomic Prediction

Once estimates of marker effects were obtained from the training data set (posterior means from the MCMC chain), the estimated BV of individual k ($\mathbf{GEBV}_k$) in the validation data set (generation 54 of population B_{val}) was computed as

$$GEBV_k = \sum_{i=1}^{m} x_{ik}\hat{g}_i,$$

where x_{ik} and $\hat{g}_i$ are the genotype and the estimated effect of genotype at locus i, respectively, and m is the total number of markers. Accuracy was calculated as the correlation between the estimated and true BV of individuals in the validation data. This accuracy was used to compare performance of the different scenarios and training populations. All scenarios were replicated 160 times, and results were averaged across replicates. Mean accuracies from alternate training populations were compared by the LSD test using the JMP software package (JMP, SAS Institute Inc., Cary, NC).

LD and Between Breed Diversity

To evaluate the extent and magnitude of LD in the training populations and its impact on accuracy, LD between pairs of SNP markers were estimated using r^2 (Hill and Robertson, 1968). Only markers with a MAF $\geq$0.1 were considered in this analysis. The power to detect LD between 2 loci is minimal when at least one of them has an extreme allele frequency (Goddard et al., 2000). Further, to evaluate the persistence of LD phase across training and validation populations, the correlations of r between the 2 populations were calculated for different distances between loci (Goddard et al., 2006).

To assess and compare the decline of LD with distance in different training populations, a nonlinear regression model was fitted to the observed r^2 between marker pairs in each training population. The model used was based on the Sved (1971) equation.

$$r^2_{ij} = 1/\left(1 + 4bd_{ij}\right) + e_{ij},$$

where r^2_{ij} is the observed LD between markers i and j in the training data, b is a coefficient that describes the decline of LD with distance in the training data, d_{ij} is distance in morgans between markers i and j, and e_{ij} is the random residual that was assumed normally distributed.

The level of genetic diversity present in the simulated breeds was investigated using Wright's *F*-statistics (Wright, 1965) F_{IT}, F_{ST}, and F_{IS}, as implemented in the program Fstat (Goudet, 2001). Genotypes at 200 loci from a random sample of 100 individuals from each of the 4 simulated breeds from generation 53 were used to estimate *F*-statistics. Significance levels for the *F*-statistics and related variance components were obtained

Table 2. Average accuracy of estimated breeding values in the validation data set (pure breed B) from genomic selection with different training data sets and marker densities (number of SNP per cM)[1]

		Marker density			
Group	Training data set	5/cM	10/cM	20/cM	40/cM
1	B	0.79[a]	0.83[a]	0.84[a]	0.85[a]
1	(Adm_AB)B	0.77[ab]	0.81[ab]	0.84[a]	0.84[ab]
1	(Adm_ABCD)B	0.76[bc]	0.80[ab]	0.84[a]	0.84[ab]
2	(A+B)	0.71[e]	0.76[de]	0.80[cd]	0.79[d]
2	(AB)	0.74[cd]	0.78[cd]	0.82[b]	0.82[c]
2	$(AB)^2$	0.72[de]	0.77[d]	0.81[bc]	0.83[bc]
3	(AB)C	0.66[f]	0.75[ef]	0.77[e]	0.79[d]
3	(AB)(CD)	0.67[f]	0.72[f]	0.79[de]	0.79[d]
4	A	0.34[h]	0.45[h]	0.48[g]	0.54[f]
4	(AC)	0.43[g]	0.50[g]	0.58[f]	0.64[e]

[a–h]Values with different letters within a column are significantly different ($P < 0.05$). Based on 160 replicates.

[1]B is the purebred B; (Adm_AB)B and (Adm_ABCD)B are admixtures of 2 and 4 breeds; (A+B) is the combined_AB; (AB) is the F_1; $(AB)^2$ is the F_2; (AB)C is the 3-way cross; (AB)(CD) is the 4-way crossbred; A is purebred A; and (AC) is cross of breeds A and C.

from 20,000 permutations and from jackknife over loci (with different loci as resampling units), as provided by the Fstat program.

RESULTS AND DISCUSSION

Accuracy of Genomic Selection

Correlations between estimated and true BV for individuals in the validation data set (population B_{val}, Figure 1) for different training populations and marker densities are shown in Table 2. Training in the same breed as the validation population (B_{54}) resulted in the greatest accuracy in all cases. Accuracies tended to be less if populations other than the validation breed were used for training, but reductions in accuracy were not significant in some cases and depended on the breed composition of the training population. Based on differences in accuracies, the training populations can be divided into 4 groups: (1) purebred B and admixed populations, (2) 2-breed combined_AB and crossed populations, (3) 3- and 4-way crosses, and (4) purebred A and AC. Training in the admixed populations resulted in similar accuracies as training in the purebred B population (group 1). The largest drop in accuracy compared with group 1 was for training populations in group 4, which included no contribution from the validation breed B. Averaged over all marker densities, relative to the accuracy of training and validating in the same breed, accuracy dropped by 46% when validation was in a different breed, whereas training in crossbred AC and validating in breed B resulted in a drop in accuracy of 35%. Populations in groups 2 and 3 had accuracies intermediate to those of groups 1 and 4. Comparing accuracies for populations in groups 2 and 3, as the number of breeds contributing to the training population increased, the accuracy dropped more. Whereas training in group 2 populations resulted in a 6% decrease of accuracy, the decrease in accuracy when using the 3- and 4-way crosses for training was on average about 10%. Interestingly, within group 2, differences in accuracy were not practically significant. Comparing groups 1, 2, and 3, the 3- and 4-way crossbred training populations showed the least accuracy of prediction.

Marker Density

Accuracy generally increased with marker density (Table 2). The increase was most noticeable when training in A, AC, and in the 3-way and 4-way crosses. Thus, the effect of marker density was more pronounced when the training population had a lesser contribution of the breed that comprised the validation population (breed B in this case). Increasing marker density from 5 to 20 markers per cM improved accuracy by 35, 20, 10, and 10% for groups 4, 3, 2, and 1, respectively. However, increasing marker density from 20 to 40 per cM did not improve accuracy as much, except for group 4 training data sets, which showed an additional increase of 11% in the accuracy. Table 2 shows that with the level of LD and the amount of breed divergence simulated in this study, marker densities as low as 5 markers per cM were sufficient for accurate prediction of BV without explicitly accounting for pedigree or breed composition in the crossbred and admixed training populations, as long as the target breed contributed to the training population.

The Effect of Time Since Divergence of Breeds

Table 3 shows the effect of the number of generations of random mating after isolation (referred to as time since divergence, **TSD**) on the accuracy of GEBV. As expected, TSD did not significantly affect accuracy

Table 3. The impact of time since divergence of breeds on the accuracy of genomic selection when training in different data sets with 5 markers per cM on a 1-morgan genome[1]

	Time since divergence	
Training data set	25	50
B	0.80^{ab}	0.80^{ab}
(Adm_AB)B	0.80^{a}	0.77^{bcd}
(Adm_ABCD)B	0.79^{abc}	0.76^{de}
(A+B)	0.76^{cde}	0.71^{g}
(AB)	0.77^{bcd}	0.74^{ef}
$(AB)^2$	0.78^{abcd}	0.72^{fg}
(AB)C	0.73^{efg}	0.66^{h}
(AB)(CD)	0.75^{def}	0.67^{h}
A	0.55^{j}	0.34^{l}
(AC)	0.61^{i}	0.43^{k}

$^{a–l}$Values with different letters within and across columns are significantly different ($P < 0.05$). Based on 160 replicates.

[1]B is the purebred B; (Adm_AB)B and (Adm_ABCD)B are admixtures of 2 and 4 breeds; (A+B) is the combined_AB; (AB) is the F_1; $(AB)^2$ is the F_2; (AB)C is the 3-way cross; (AB)(CD) is the 4-way crossbred; A is purebred A; and (AC) is cross of breeds A and C.

when the purebred B population was used as training data, but accuracies significantly increased for all other training data sets when TSD was reduced from 50 to 25. The maximum increase in accuracy was observed when using group 4 populations [A or (AC)] for training to predict breed B, for which accuracies increased by 62 and 42%, respectively. Considering Tables 2 and 3, for both values of TSD, training in the admixed populations (group 1) resulted in a greater accuracy than training in the crossbred populations (groups 2 and 3).

The above results were based on one chromosome of 1 morgan to make the simulation computationally feasible. To determine if the simulated genome size affects the main conclusions, an additional data set with a total of 5,000 markers and 100 QTL on 5 chromosomes, each of length 1 morgan, was simulated. The analysis was run for the training populations of purebred B, [(Adm_ABCD)B], combined_AB, and 4-way crossbred only. The corresponding accuracies for this scenario were 0.80, 0.77, 0.72, and 0.71, respectively. These values are comparable with the accuracies presented in the Table 2 for the scenario of 10 markers per cM.

Extent of LD and Differentiation Between Breeds

To explain the differences in accuracy between the groups described earlier and shown in Table 2, LD in the different training populations was examined by comparing the average distances between flanking markers at different levels of r^2 (Table 4). There were significant differences in the extent of LD between the training populations. Figure 2 depicts how LD decayed with distance in different training populations and shows significant differences in the rate of breakdown of LD between the populations. Note that for all training data sets, average LD was between the expected LD based on Sved's (1971) formula for effective population sizes of 100 and 500. The slowest and the steepest rates of decline of LD were in the purebred and the 4-way crossbred training populations, respectively. The combined_AB and the 2-way crosses had a slower rate of decay of LD than the 3-way crosses. Nonlinear regression was used to estimate a coefficient that describes the rate of decay of LD with distance in each training population, based on the Sved (1971) formula. Resulting estimates of rate of decay of LD are shown in Table 4. In Sved (1971), the rate of decay constant was an estimate of the effective population size, but this assumes a closed random mating population with constant historical effective population size. The interpretation is not valid for the populations analyzed here because effective population size was not constant for the purebred population and the other populations also represented crosses and admixtures. Nevertheless, the estimated constants give a good indication that the effective number of founders and, therefore, the extent and decline of LD with distance, differed substantially between populations.

Wright's *F*-statistics were used to quantify the amount of divergence between the simulated breeds in generation 53. The estimated F_{IT}, F_{ST}, and F_{IS} were 0.240 (SE = 0.011), 0.236 (SE = 0.011), and 0.005 (SE = 0.004), respectively, when breed separation was in generation 0.

Training data sets consisting of a purebred, a 2-breed combined, several crossbreds, and admixed populations were compared for their ability to accurately predict true BV of selection candidates in a purebred population using GS. In the following, the main focus will be on crossbred, combined, and admixed populations.

Accuracy of Genomic Selection

In Table 2, all types of training populations performed remarkably well, except when another pure breed was used for training or when the training data consisted of a cross that did not include the target breed (AC). The admixed training populations resulted in nearly the same accuracies as when training was in the breed to which selection candidates belonged (breed B).

Extent of LD

Based on results presented in Table 4, when considering the average distance between pairs of adjacent markers that had $r^2 \geq 0.1$, there was more extensive LD in the crossbred, combined_AB, and admixed populations than in the purebred population. The extent of LD is proportional to the age of the LD generating event (Reich et al., 2001). In a subdivided or crossbred population, like the combined_AB and the F_1 training populations, LD is composed of 2 parts (Nei and Li,

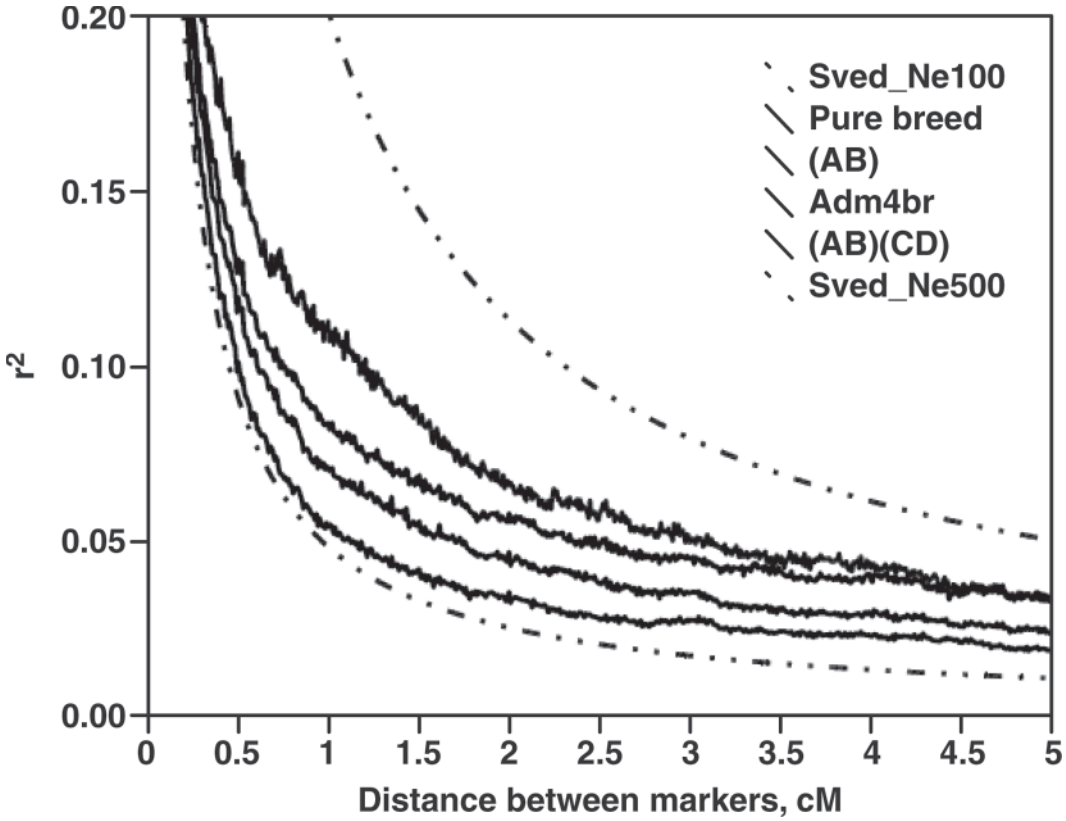

Figure 2. Average linkage disequilibrium as measured by of r^2 against distance (cM) in different training populations. Sved_Ne100 and Sved_Ne500 are expectations based on Sved (1971) $E(r^2) \approx 1/(1 + 4N_e c)$, where $N_e = 100$ or 500 and c is recombination rate calculated as $0.5(1 - \exp(-2 \times \text{map distance})$. Pure breed is the purebred B training population; (AB) is the F_1; Adm4br is the admixture of 4 breeds, and (AB)(CD) is the 4-way crossbred training population. The graph is based on 60 replicates; average r^2 over all replicates are plotted against distance. Lines in the graph are in the order as shown in the legend. The top line is Sved_Ne100, second line is purebred, third line is (AB), fourth line is Adm4br, fifth line is (AB)(CD), and the last line is Sved_Ne500. Other training populations are not shown for clarity of the picture.

1973; Lo et al., 1993). The first part is the average LD that existed within the parental populations, referred to as old LD, and the second part is LD generated in the cross as a result of difference in gene frequencies between the parental breeds, referred to as new LD because it is created by a recent phenomenon. Whereas the old LD is confined to shorter distances due to the accumulation of recombination events, the new LD extends over longer intervals. The combined_AB training population was composed of breeds A and B with equal proportion; as a result, the distribution of LD in this population was the same as in the F_1 training population. Figure 3 depicts the average distance between pairs of markers at various levels of LD.

Table 4. Average distance (in cM) between adjacent markers with r^2 greater than 0.1, 0.4, or 0.7 in different training data sets, with the percentage of such marker pairs out of all adjacent pairs with r^2 greater than 0 in parentheses[1,2,3,4]

	Training population						
Item	B	(AB)	(A+B)	$(AB)^2$	(AB)C	Adm4br	(AB)(CD)
Beta estimate	151	262	263	269	349	355	440
Minimum r^2							
0.7	0.31^{a} (0.42)	0.21^{b} (0.21)	0.21^{b} (0.21)	0.16^{c} (0.22)	0.12^{d} (0.17)	0.10^{e} (0.17)	0.09^{f} (0.14)
0.4	0.75^{c} (1.0)	2.21^{a} (0.58)	2.17^{b} (0.58)	0.50^{d} (0.55)	0.56^{e} (0.41)	0.32^{f} (0.41)	0.21^{g} (0.33)
0.1	3.22^{g} (5.7)	15.12^{a} (7.2)	15.05^{b} (7.2)	4.34^{e} (4.4)	10.5^{c} (4.1)	7.39^{d} (3.4)	4.41^{f} (2.3)

[a–g]Values with different letters within each row are significantly different ($P < 0.05$).

[1]Estimated coefficients of linkage disequilibrium decline (Beta estimate) are shown in the first row of the table.

[2]B is the purebred B training population; (AB) is the F_1; (A+B) is the combined_AB, $(AB)^2$ is the F_2; (AB)C is the 3-way cross, Adm4br is the admixture of 4 breeds, and (AB)(CD) is the 4-way crossbred training population.

[3]Based on 60 replicates; in each replicate, distances were averaged across adjacent pairs that met the minimum r^2 value, resulting in at least 100,000 pairs per replicate).

[4]All Beta estimates had SE of less than 1.

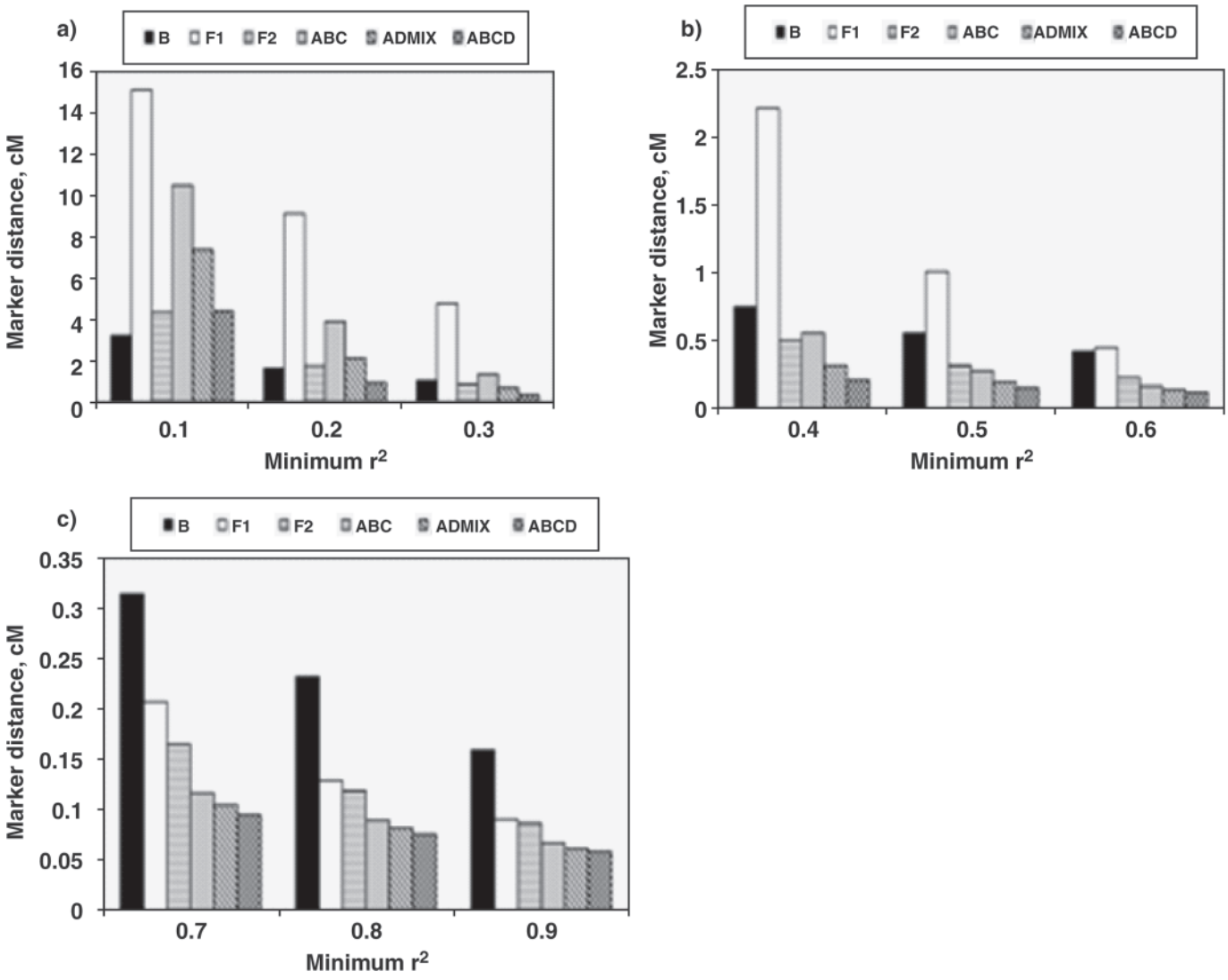

Figure 3. Average distance (cM) between adjacent markers in different training populations at various levels of linkage disequilibrium (LD; r^2). B is the purebred B, F1 is the (AB), F2 is the $(AB)^2$, ABC is the 3-way crossbred, ADMIX is the admixture of 4 breeds, and ABCD is the 4-way cross training population. Based on 60 replicates, results were averaged over distances with certain amount of LD and over replicates. Note the different scales of the graphs. a) Minimum r^2 between 0.10 and 0.30, b) minimum r^2 between 0.40 and 0.60, c) minimum r^2 between 0.70 and 0.90.

Table 4, on the other hand, shows that haplotype with strong LD ($r^2 \geq 0.70$) are significantly shorter in the admixed and the crossbred populations compared with the purebred population. The average distance between pairs of markers with strong LD is 3 times larger in the purebred than in the admixed and 4-way crossbred populations (Table 4 and Figure 3c). This very narrow region of strong LD in the crossbred training populations might explain the high accuracy obtained with these populations (Table 2). In the same way that LD limited to short distances is beneficial in QTL fine mapping by providing a more accurate estimate of the QTL position (Pritchard and Przeworski, 2001; Reich et al., 2001; Aerts et al., 2007), it can also result in greater accuracy of GS because only markers that are very close to the QTL will explain a high proportion of the QTL variance and this association will not rapidly erode over generations by recombination.

In a human genetics study, Shifman and Darvasi (2001) compared the average level of LD between SNP markers for distances below and over 200 kb in an outbred population (Centre d'Etude du Polymorphisme Humain) with that in several isolated populations (Finnish, Ashkenazi, and Sardinian). Their findings showed that at short intervals the amount of LD in the outbred population was comparable with that in the isolated populations, whereas at long intervals (>200 kb) there was up to 6 times more LD in the isolated populations than that in the outbred population. In another study, Shifman et al. (2003) compared an admixed, an outbred, and an isolated population (African Americans, Caucasians, and Ashkenazi Jews, respectively). They found that the average LD declined with distance between loci more rapidly in the admixed population. This is in accordance with our results, which also showed that the average level of LD was

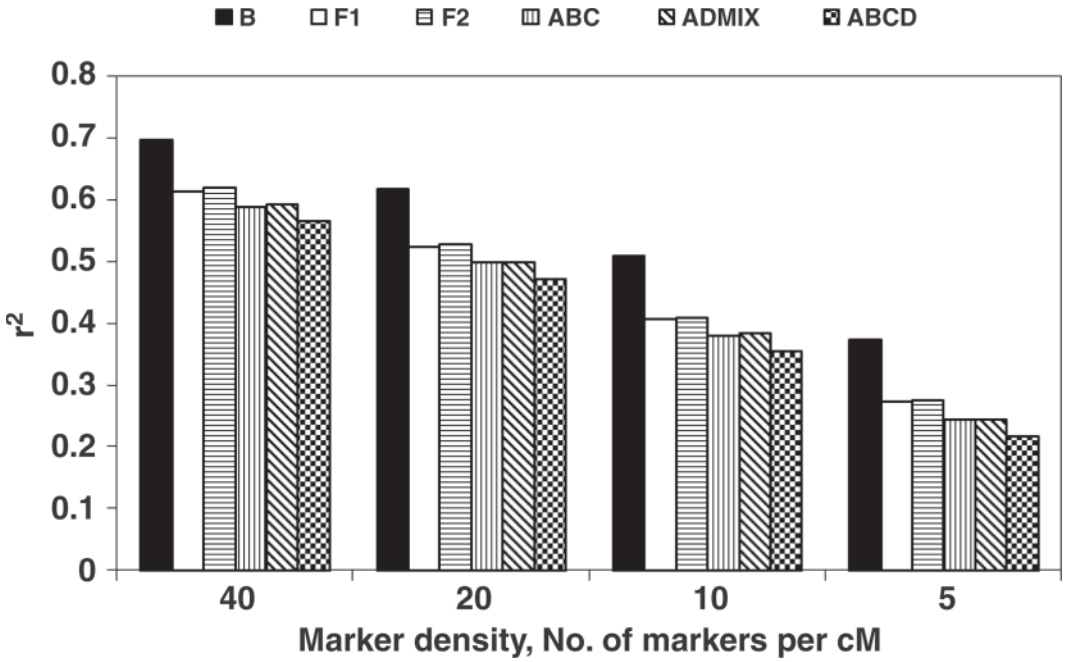

Figure 4. The average level of linkage disequilibrium as a function of marker density (No. of markers per cM) and type of training population. B is the purebred B, F1 is the (AB), F2 is the $(AB)^2$, ABC is the 3-way crossbred, ADMIX is the admixture of 4 breeds, and ABCD is the 4-way crossbred training population. Other training populations are not shown, because they showed the same trend and magnitude of linkage disequilibrium. The graph is based on 60 replicates; for each training population, average r^2 of all replicates for marker distances of 0.025, 0.05, 0.10, and 0.20 cM was calculated.

greater and extended over longer intervals in the purebred population compared with crossbred and admixed populations (see Figures 2, 3c, and 4). In a crossbred population, individuals are more distantly related to each other (i.e., the mean time to a common ancestor is longer); thus, LD haplotypes in the population are narrower than those in a purebred population. Results of the canine genome project have also shown that haplotype blocks are several Mb long within a breed, but they are much shorter across breeds, extending only to tens of kb (Lindblad-Toh et al., 2005). The ancient domestic dog diverged from wolves 15,000 to 100,000 yr ago, whereas most of the new breeds of dog were formed within the past few hundred years (Lindblad-Toh et al., 2005).

Figure 2 illustrates that LD in all training populations fell between the expected LD based on the Sved (1971) formula for effective population sizes of 100 and 500. As can be seen from the figure, LD at short distances followed the expectation based on $N_e = 500$, whereas at larger distances it tended toward its expectation based on $N_e = 100$. The LD at short intervals is a function of N_e in the distant past, whereas LD at longer intervals reflects N_e in the recent past (Hayes et al., 2003).

Marker Density

As it is evident from Table 4, the frequency with which strong LD haplotypes occur differs substantially between the training populations. Consider the difference of accuracies (Table 2) when the training population is purebred B or a 4-way crossbred population, for example. Whereas the high LD signals are restricted to very short distances in the 4-way crossbred population, there are about 3 times as many markers with strong LD in the purebred B population (Table 4). This might describe why the 3- and 4-way crossbred populations were much more affected by marker density compared with the purebred training population. Increasing marker density is expected to raise the level of LD between markers and QTL because the average distance between adjacent loci is inversely related to marker density and recombination is less likely to erode associations between tightly linked loci. Figure 4 illustrates the relationship between marker density and the level of LD between markers. The greater the LD between a pair of loci (in fact a marker and QTL), the larger is the variance that is associated with the marker (Luo et al., 1997).

The effect of marker density on the accuracy of GS has been discussed in some recent studies (Muir, 2007; Calus et al., 2008; Solberg et al., 2008). Solberg et al. (2008) simulated SNP markers with several densities. In their study, in which training and validation populations were purebreds, accuracies of GS using SNP markers with densities of 1, 2, 4, and 8 markers per cM were found 0.69, 0.79, 0.84 and 0.86, respectively (with 1,000 QTL on the genome). For the density of 4 SNP per cM, their accuracy of 0.84 is roughly comparable with our estimate of 0.79 (Table 2), keeping in mind that we had 5 SNP per cM. The greater heritability of the trait (0.5 vs. 0.30) considered in Solberg et al. (2008) might be a reason for the difference in accuracies obtained in

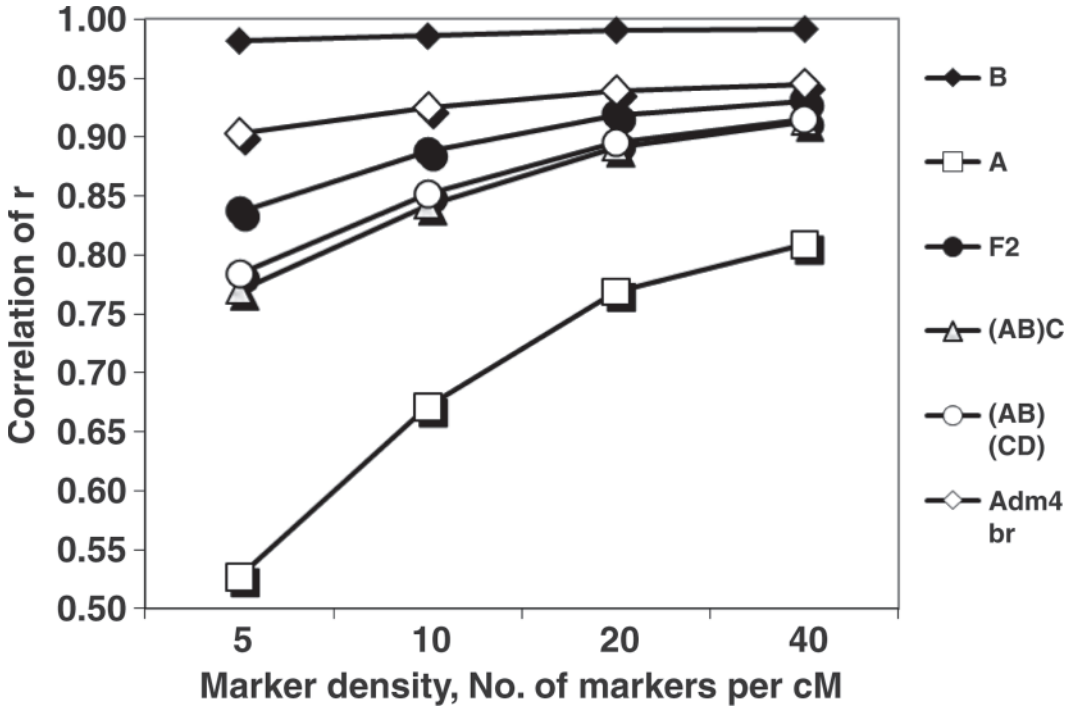

Figure 5. Correlation of r between each pair of training and validation populations, as a function of marker density (in 1 cM). r is the correlation coefficient; B is training and validating in breed B; A is training in breed A and validating in breed B; F2 is training in $(AB)^2$ and validating in B; (AB)C is training in the 3-way cross and validating in B; (AB)(CD) is training in a 4-way cross and validating in B; and Adm4Br is training in the admixture of 4 breeds and validating in breed B. Based on 60 replicates.

the 2 studies. With a reduced heritability, markers explain a smaller proportion of additive genetic variance and decreased accuracy will be obtained (Goddard and Hayes, 2007). Thus, to get accurate estimates of marker effects, a larger sample size is required. The level of LD between markers and QTL and the sample size used to estimate the QTL effects are the 2 factors driving the accuracy of marker assisted selection and the power of QTL detection (Lande and Thompson, 1990; Hayes et al., 2007).

Persistence of LD Phase

Markers in LD with putative QTL are valuable for marker-assisted selection if the marker-QTL linkage-phase and extent of LD is consistent between the population used for estimation and the population in which selection is to be practiced (Dekkers and Hospital, 2002; Goddard et al., 2006). Figure 5 shows the persistence of LD phase between adjacent markers in the training and validation populations, as measured by the correlation of r between the 2 populations. A greater correlation implies that the marker-marker (and most probably the marker-QTL) linkage phase is more consistent between the 2 populations. This figure shows that the correlation of r increased with marker density and was less if the training and validation populations were more different (e.g., when training and validation was in different breeds vs. in the same breed). This relationship between persistence of LD phase and divergence between breeds agrees with other reports (e.g., Andreescu et al., 2007; Gautier et al., 2007; de Roos et al., 2008). Obviously, the shorter the length of a haplotype, the greater is the chance of its similarity across populations. In the same way, as distance in time between 2 subpopulations increases, there is a greater chance for recombination to break down the LD that was present in the ancestral population and drift to create new LD within each subpopulation (Hill and Robertson, 1968, Goddard et al., 2006).

In a study of extent and persistence of LD phase in Holstein-Friesian, Jersey, and Angus cattle, de Roos et al. (2008) reported a correlation close to 1 of r between 2 breeds for pairs of markers that were <10 kb apart and a decline of this correlation as distance between markers or divergence between breeds increased. Considering marker loci that were less than 10 kb apart, Gautier et al. (2007) reported correlations of 0.54 to 0.93 (0.77, on the average) of r for pairs of European cattle breeds, which again reflects how the degree of relationship between 2 breeds changes the correlation of r. In our simulations, marker densities were not greater than 40 per cM (a distance of ≈25 kb between adjacent markers). For pairs of markers with such average distance, the correlation of r for across breeds GS was 0.81 (see Figure 5). For the LD correlation between 2 breeds to be high, tight LD between a pair of loci should exist in the ancestral population before divergence of the 2 breeds such that recombination cannot erode it (Goddard et al., 2006).

The high correlation of r for the admixed training populations compared with the crossbred training populations reveals a closer relationship between these populations and breed B because the admixed populations had a greater proportion of breed B genes. Correlation of r between 2 populations can be used as a good estimate of relationship between the 2 populations (Andreescu et al., 2007). This could explain why the admixed populations resulted in greater accuracy of selection than the crossbred training populations. The greater accuracy obtained when training in the AC population vs. in the A purebred population (Table 2) might be explained as follows. The use of AC cross forces the model to look only at ancestral LD that was already present at the time of separation of breeds, rather than using the new LD. Ancestral LD is more likely to be present in breed B as well. This explains why the correlation of r between the AC and breed B was greater than the correlation of r between breeds A and B (data not shown). The correlation of r between 2 populations may be used as an indication of the required marker density to ensure marker-QTL linkage-phase persists across the populations (Goddard et al., 2006).

In a simulation study, Ibanez-Escriche et al. (2009) applied GS to select purebreds for crossbred performance. In their work, they compared the performance of a model with breed-specific effects to a model with the same effects across breeds. It was shown that for 2 unrelated breeds, where correlation of r between the 2 breeds was 0, the across-breed model was as accurate as the breed-specific model in prediction of BV. The breed-specific model resulted in decreased accuracy of prediction when the marker density increased compared with the across-breed model (Ibanez-Escriche et al., 2009). More effects might need to be estimated in a multi-breed population. When alternative alleles are fixed in different breeds, there are almost twice as many effects to be estimated compared with the purebred population. Thus, one may need a larger sample size in a multi-breed population to get an accuracy that is comparable with the accuracy in a purebred population.

Divergence of the Breeds

Wright's *F*-statistics are inbreeding coefficients that differ in the reference population that is used (Hartl and Clark, 1989). The F_{IT} is the broadest measure of inbreeding in that it takes into account the effects of nonrandom mating within the subpopulations (F_{IS}) and the effects of population subdivision (F_{ST}; Hartl and Clark, 1989). The estimate of F_{IS} of 0.005 implies only a minor deficit of heterozygosity within breeds. Because individuals in each breed were randomly mated, a significant divergence from the Hardy-Weinberg proportions within each breed is not expected. The expected value of F_{IS} is $1/(2N_e)$, which with $N_e = 100$ in each breed agrees with the results and indicates negligible levels of inbreeding within the breeds. The expected value of F_{ST} under the conditions of an idealized population with subdivision (Falconer, 1989) is $1 - (1 - 1/2N_e)^t$, which with $t = 53$ generations is equal to 0.233 and is in close agreement with our estimate based on marker data of 0.236. This value of F_{ST} shows that about 24% of the total genetic variability in the whole population can be attributed to the difference among breeds [e.g., Cañón et al. (2001)], or that about 24% of shared allelic diversity was lost within each breed since they were separated. Thus, the breeds had significantly diverged from each other. With $F_{IS} = 0$, F_{ST} and F_{IT} are expected to be equal because $(1 - F_{IS})(1 - F_{ST}) = (1 - F_{IT})$ (Wright, 1969). Recently, McKay et al. (2008) published estimated pairwise and global F_{ST} values for several cattle breeds, based on a panel of 2641 SNP. Their estimated global F_{ST} when they considered both *Bos taurus* and *Bos indicus* breeds was 0.29. However, the estimated global F_{ST} reduced to 0.17 when they excluded the *Bos indicus* breeds from their analysis (McKay et al., 2008). Therefore, our simulated breeds had enough divergence to represent current breeds of beef cattle.

The Effect of TSD of Breeds

The accuracies for the scenarios of TSD = 25 and 50 (Table 3) were compared. As expected, no effect of TSD on accuracy was observed for the scenario of training and validating in the same breed (B) because there was no divergence within the same breed. The minimum (4%) and the maximum (62%) increase in accuracies were observed for the admixed training populations and when training and validating in different breeds, respectively, when TSD changed from 50 to 25 generations. Again, this reflects the fact that the more distantly 2 populations are related, the greater is the chance of recombination to break down the shared ancestral haplotypes (and even reverse the LD phase) across the populations. This might explain why accuracy of GS was reduced for training populations other than breed B when TSD changed from 25 to 50 (Table 3). The more time elapsed since separation of 2 subpopulations, the greater is the loss of shared allelic diversity between them (McKeigue, 2005).

Effect of Selection

In this simulation only LD generated by mutation and drift was considered. In reality, livestock populations have been under selection for a long time and breeds may have been under varying intensities and directions of selection. To assess the impact of differential selection of breeds on the validity of our results, the relationship, across replicates, between the accuracy of GS and the difference in the mean true BV of the breeds that were

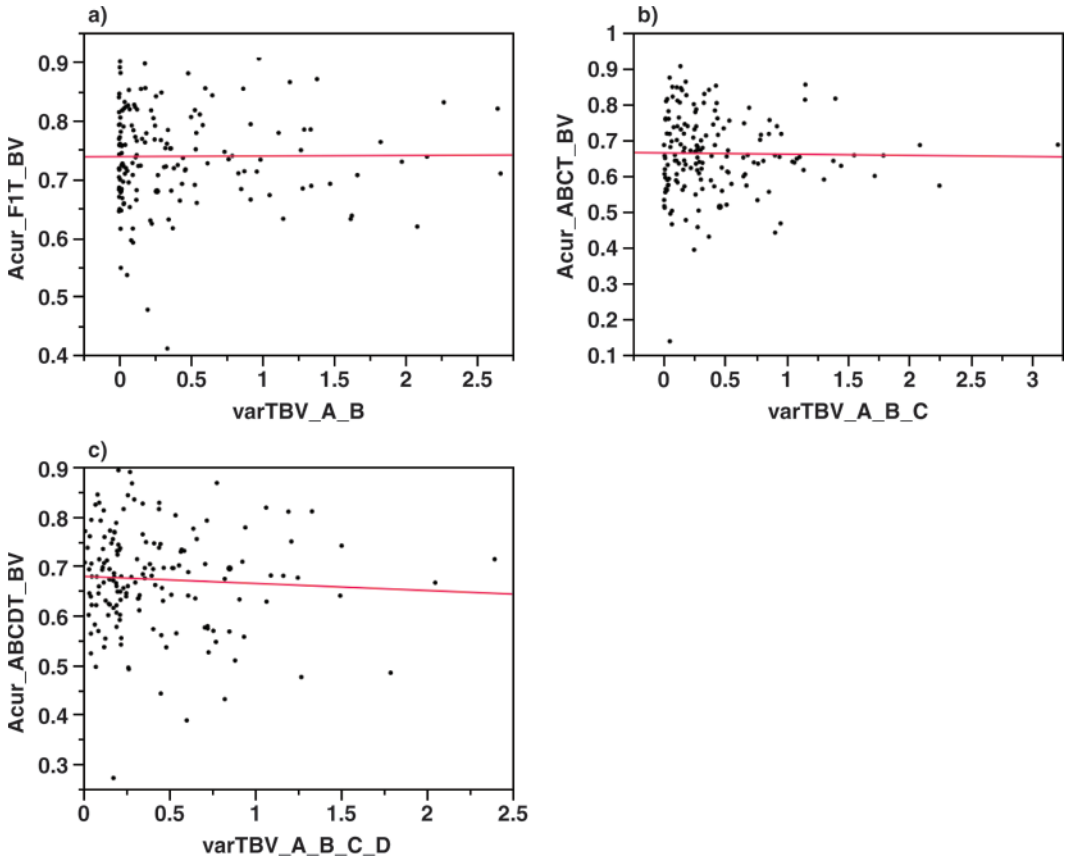

Figure 6. Plot of accuracy (Acur) against between breeds variance of true breeding values (varTBV). a) Plot of accuracy when training in an F_1 vs. between breeds [A and B] variance of true breeding values; b) plot of accuracy when training in a 3-way cross [(AB)C] vs. between breeds (A, B, and C) variance of true breeding values; c) plot of accuracy when training in a 4-way cross [(AB)(CD)] vs. between breeds (A, B, C, and D) variance of true breeding values. In all plots, the black line shows the regression of accuracy on between breeds variance of true breeding values.

crossed or admixed was evaluated. Although any breed differences in the simulations were the result of mutation and drift, rather than directed selection, selection can be viewed as directed random drift. Therefore, the accuracy of GS against the variance of the mean true BV of the breeds that are crossed was plotted for the 160 replicates of our simulation (Figure 6). This did not reveal any significant association of accuracy with the extent of the diversity of breeds contributing to the cross. Considering Figure 6a, for example, although breeds A and B showed quite a range of different true genetic means for the trait of interest, this difference did not affect the accuracy of selection.

Traditionally, GS studies by simulation have only considered additive QTL effects (Meuwissen et al., 2001). However, in reality QTL contribute to total genetic variation by themselves or by interacting with other QTL (Carlborg et al., 2006). Interaction among loci might result in a biased estimate of the effect of each locus (Carlborg and Halley, 2004). In a recent study, Carlborg et al. (2006) identified a genetic network of several interacting loci that significantly contributed to BW at 56 d of age in chickens. Their results showed that the power of QTL mapping experiment in identifying loci whose effect is dependent on the genotype at another locus improves when the inter- and intralocus interactions are included in their statistical model. Thus, dominance and epistatic QTL effects might need to be considered in GS studies where the objective is improvement of purebreds for their crossbred progeny performance.

In this study our focus was on purely additive gene effects; however, there might be the question of how accurate GS predictions will be in the presence of heterosis. In an F_1 population, where the population is homogenous in terms of breed composition, we think ignoring the effects of heterosis does not bias the prediction of marker effects. However, in an admixed population, because individuals have different breed compositions, the dominance effects must be explicitly accounted for.

Another question might be the choice of training population when the selection candidates are crossbred themselves. In a recent simulation study, Ødegård et al. (2009) investigated introgression of favorable alleles from an inferior donor line into a superior recipient line using dense marker genotyping and GS. Their proposed method of combining backcrossing and GS increased the frequency of favorable QTL alleles at the expense of unfavorable ones (irrespective of origin) across the entire genome, without any specific effort to reduce the linkage drag from the donor line (Ødegard et al., 2009).

In a recent study with real data, Harris et al. (2008) compared the accuracy of GS of purebred Jersey (**J**), purebred Holstein-Friesian (**HF**), and crossbred J-HF bulls using the BovineSNP50 BeadChip. The training data sets were either 1 of the 2 breeds (J or HF) or a combined data set of both breeds. Training in one breed and validating in another breed resulted in an accuracy of −0.10 to 0.3. Accuracy of GS of crossbred J-HF bulls was 5 to 10% greater when training was done in the combined data set compared with when training was in J or HF breeds (Harris et al., 2008). Assuming that the validation population is a crossbred, we compared the accuracy of GS when the training population was purebred or crossbred (F_1). Training in the crossbred population increased accuracy of GS in the crossbred population by 11% compared with training in the purebred (data not shown).

A population that is a crossbred or an admixture of different breeds can be used as a training data set for GS and can provide reasonably accurate estimates of true BV of purebred selection candidates. This also implies that, with GS using high-density SNP markers, marker estimates obtained from crossbred populations can be used to select purebreds for crossbred performance, as suggested by Dekkers (2007), and examined by Ibanez-Escriche et al. (2009). Our results showed that in crossbred and admixed populations, haplotypes with strong LD are much shorter than in purebred populations. Thus, crossbred or admixed populations are more suitable for QTL fine mapping than purebred populations, provided marker density is sufficient.

Furthermore, because haplotype segments with strong LD in crossbred and admixed populations are narrower, markers in such segments are expected to have more consistent associations with QTL across the training and validation populations. Therefore, the decline of accuracy of GS over generations that has been observed in simulation studies (e.g., Habier et al., 2007) might be slower when admixed or crossbred populations are used for training than when purebred populations are used. By combining 2 pure breeds into a single training population, one can take advantage of a larger sample size for simultaneous estimation of marker effects and thus improve the accuracy of GS. In our simulation, when the size of the training population for the combined_AB training population was doubled, a 7% increase of the accuracy resulted (data not shown). In addition, by combining breeds into a single training population (vs. making certain crosses like an F_1), a lot of time and effort can be saved. More importantly, there is a greater chance of segregation of breed-specific QTL in a multibreed training population.

In the present study, while dealing with admixed populations, the population structure or additive genetic relationships were not explicitly modeled, which might be regarded as the standard method to limit the false discoveries due to population admixture in marker-phenotype association studies. Nevertheless, GS using high-density markers proved to be efficient enough to distinguish between true signals of association from spurious signals, at least under the idealized population structures that were used in the simulations. Whether or not this could provide an alternative methodology for association studies in populations with cryptic structures or extensive genealogical relationships requires further research.

LITERATURE CITED

Aerts, J., H. J. Megens, T. Veenendaal, I. Ovcharenko, R. Crooijmans, L. Gordon, L. Stubbs, and M. Groenen. 2007. Extent of linkage disequilibrium in chicken. Cytogenet. Genome Res. 117:338–345.

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont, and J. C. M. Dekkers. 2007. Linkage disequilibrium in related breeding lines of chickens. Genetics 177:2161–2169.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561.

Cañón, J., P. Alexandrino, I. Bessa, C. Carleos, Y. Carretero, S. Dunner, N. Ferran, D. Garcia, J. Jordana, D. Laloë, A. Pereira, A. Sanchez, and K. Moazami-Goudarzi. 2001. Genetic diversity measures of local European beef cattle breeds for conservation purposes. Genet. Sel. Evol. 33:311–332.

Carlborg, Ö., and C. S. Haley. 2004. Epistasis: Too often neglected in complex trait studies? Nat. Rev. Genet. 5:618–625.

Carlborg, Ö., L. Jacobsson, P. Åhgren, P. Siegel, and L. Andersson. 2006. Epistasis and the release of genetic variation during long-term selection. Nat. Genet. 38:418–420.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. Genetics 179:1503–1512.

Dekkers, J. C., and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3:22–32.

Dekkers, J. C. M. 2007. Marker-assisted selection for commercial crossbred performance. J. Anim. Sci. 85:2104–2114.

Falconer, D. S. 1989. Introduction to Quantitative Genetics. 3rd ed. Longman Wiley, Burnt Mill, Harlow, Essex, UK.

Flint-Garcia, S. A., J. M. Thornsberry, and E. S. t. Buckler. 2003. Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54:357–374.

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, C. Grohs, A. Boland, J.-G. Garnier, D. Boichard, G. M. Lathrop, I. G. Gut, and A. Eggen. 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177:1059–1070.

Goddard, K. A. B., P. J. Hopkins, J. M. Hall, and J. S. Witte. 2000. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am. J. Hum. Genet. 66:216–234.

Goddard, M. E., and B. J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124:323–330.

Goddard, M. E., B. J. Hayes, H. McPartlan, and A. J. Chamberlain. 2006. Can the same genetic markers be used in multiple breeds? Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brazil, August 13–18, 2006. CD-ROM Commun. No. 22-16.

Goudet, J. 2001. FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). http://www.unil.ch/izea/softwares/fstat.html Accessed Nov. 25, 2008.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Harris, B. L., D. L. Johnson, and R. J. Spelman. 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. Pages 325–330 in Proc. Interbull Meet., June 16–19, 2008, Niagara Falls, NY.

Hartl, D. L., and A. G. Clark. 1989. Principles of Population Genetics. 2nd ed. Sinauer, Sunderland, MA.

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genet. Res. 89:215–220.

Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard. 2003. Novel multi-locus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13:635–643.

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

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231.

Hirschhorn, J. N., and M. J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6:95–108.

Ibanez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41:12–29.

Karlin, S. 1984. Theoretical aspects of genetic map functions in recombination processes. Pages 209–228 in Human Population Genetics: The Pittsburgh Symposium. A. Chakravarti, ed. Van Nostrand Reinhold, New York, NY.

Kennedy, B. W., M. Quinton, and J. A. van Arendonk. 1992. Estimation of effects of single genes on quantitative traits. J. Anim. Sci. 70:2000–2012.

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

Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, M. Kamal, M. Clamp, J. L. Chang, E. J. Kulbokas, M. C. Zody, E. Mauceli, X. Xie, M. Breen, R. K. Wayne, E. A. Ostrander, C. P. Ponting, F. Galibert, D. R. Smith, P. J. Dejong, E. Kirkness, P. Alvarez, T. Biagi, W. Brockman, J. Butler, C. W. Chin, A. Cook, J. Cuff, M. J. Daly, D. Decaprio, S. Gnerre, M. Grabherr, M. Kellis, M. Kleber, C. Bardeleben, L. Goodstadt, A. Heger, C. Hitte, L. Kim, K. P. Koepfli, H. G. Parker, J. P. Pollinger, S. M. Searle, N. B. Sutter, R. Thomas, C. Webber, J. Baldwin, A. Abebe, A. Abouelleil, L. Aftuck, M. Ait-Zahra, T. Aldredge, N. Allen, P. An, S. Anderson, C. Antoine, H. Arachchi, A. Aslam, L. Ayotte, P. Bachantsang, A. Barry, T. Bayul, M. Benamara, A. Berlin, D. Bessette, B. Blitshteyn, T. Bloom, J. Blye, L. Boguslavskiy, C. Bonnet, B. Boukhgalter, A. Brown, P. Cahill, N. Calixte, J. Camarata, Y. Cheshatsang, J. Chu, M. Citroen, A. Collymore, P. Cooke, T. Dawoe, R. Daza, K. Decktor, S. Degray, N. Dhargay, K. Dooley, K. Dooley, P. Dorje, K. Dorjee, L. Dorris, N. Duffey, A. Dupes, O. Egbiremolen, R. Elong, J. Falk, A. Farina, S. Faro, D. Ferguson, P. Ferreira, S. Fisher, M. Fitzgerald, K. Foley, C. Foley, A. Franke, D. Friedrich, D. Gage, M. Garber, G. Gearin, G. Giannoukos, T. Goode, A. Goyette, J. Graham, E. Grandbois, K. Gyaltsen, N. Hafez, D. Hagopian, B. Hagos, J. Hall, C. Healy, R. Hegarty, T. Honan, A. Horn, N. Houde, L. Hughes, L. Hunnicutt, M. Husby, B. Jester, C. Jones, A. Kamat, B. Kanga, C. Kells, D. Khazanovich, A. C. Kieu, P. Kisner, M. Kumar, K. Lance, T. Landers, M. Lara, W. Lee, J. P. Leger, N. Lennon, L. Leuper, S. LeVine, J. Liu, X. Liu, Y. Lokyitsang, T. Lokyitsang, A. Lui, J. Macdonald, J. Major, R. Marabella, K. Maru, C. Matthews, S. McDonough, T. Mehta, J. Meldrim, A. Melnikov, L. Meneus, A. Mihalev, T. Mihova, K. Miller, R. Mittelman, V. Mlenga, L. Mulrain, G. Munson, A. Navidi, J. Naylor, T. Nguyen, N. Nguyen, C. Nguyen, T. Nguyen, R. Nicol, N. Norbu, C. Norbu, N. Novod, T. Nyima, P. Olandt, B. O'Neill, K. O'Neill, S. Osman, L. Oyono, C. Patti, D. Perrin, P. Phunkhang, F. Pierre, M. Priest, A. Rachupka, S. Raghuraman, R. Rameau, V. Ray, C. Raymond, F. Rege, C. Rise, J. Rogers, P. Rogov, J. Sahalie, S. Settipalli, T. Sharpe, T. Shea, M. Sheehan, N. Sherpa, J. Shi, D. Shih, J. Sloan, C. Smith, T. Sparrow, J. Stalker, N. Stange-Thomann, S. Stavropoulos, C. Stone, S. Stone, S. Sykes, P. Tchuinga, P. Tenzing, S. Tesfaye, D. Thoulutsang, Y. Thoulutsang, K. Topham, I. Topping, T. Tsamla, H. Vassiliev, V. Venkataraman, A. Vo, T. Wangchuk, T. Wangdi, M. Weiand, J. Wilkinson, A. Wilson, S. Yadav, S. Yang, X. Yang, G. Young, Q. Yu, J. Zainoun, L. Zembek, A. Zimmer, and E. S. Lander. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

Lo, L. L., R. L. Fernando, and M. Grossman. 1993. Covariance between relatives in multibreed populations: Additive model. Theor. Appl. Genet. 87:423–430.

Luo, Z. W., R. Thompson, and J. A. Woolliams. 1997. A population genetics model of marker-assisted selection. Genetics 146:1173–1183.

McKay, S. D., R. D. Schnabel, B. M. Murdoch, L. K. Matukumalli, J. Aerts, W. Coppieters, D. Crews, E. Dias Neto, C. A. Gill, C. Gao, H. Mannen, Z. Wang, C. P. Van Tassell, J. L. Williams, J. F. Taylor, and S. S. Moore. 2008. An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. BMC Genet. 9:37–45.

McKeigue, P. M. 2005. Prospects for admixture mapping of complex traits. Am. J. Hum. Genet. 76:1–7.

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

Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard. 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161:373–379.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124:342–355.

Nei, M., and W. H. Li. 1973. Linkage disequilibrium in subdivided populations. Genetics 75:213–219.

Ødegård, J., H. Yazdi, A. Sonesson, and T. H. Meuwissen. 2009. Incorporating desirable genetic characteristics from an inferior into a superior population using genomic selection. Genetics 181:737–745.

Plummer, M., N. Best, K. Cowles, and A. K. Vines. 2006. Coda: Convergence diagnosis and output analysis for MCMC. R News 6:7–11.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909.

Pritchard, J. K., and M. Przeworski. 2001. Linkage disequilibrium in humans: Models and data. Am. J. Hum. Genet. 69:1–14.

Pritchard, J. K., M. Stephens, N. A. Rosenberg, and P. Donnelly. 2000. Association mapping in structured populations. Am. J. Hum. Genet. 67:170–181.

Rabinowitz, D. 1997. A transmission disequilibrium test for quantitative trait loci. Hum. Hered. 47:342–350.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Shifman, S., and A. Darvasi. 2001. The value of isolated populations. Nat. Genet. 28:309–310.

Shifman, S., J. Kuypers, M. Kokoris, B. Yakir, and A. Darvasi. 2003. Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. 12:771–776.

Solberg, T., A. Sonesson, J. Woolliams, and T. Meuwissen. 2008. Genomic selection using different marker types and densities. J. Anim. Sci. 86:2447–2454.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52:506–516.

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

Wang, W. Y. S., B. J. Barrat, D. G. Clayton, and J. A. Todd. 2005. Genome-wide association studies: Theoretical and practical concerns. Nat. Rev. Genet. 6:109–118.

Wright, S. 1965. The interpretation of population structure by F-statitics with special regard to systems of mating. Evolution 19:395–420.

Wright, S. 1969. Evolution and the genetics of populations. Page 295 in Vol. 2: The Theory of Gene Frequencies. University of Chicago Press, Chicago, IL.

Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. Genetics 163:789–801.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208.