Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes

K. Kizilkaya, R. L. Fernando and D. J. Garrick

J ANIM SCI 2010, 88:544-551.
doi: 10.2527/jas.2009-2064 originally published online October 9, 2009

Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes[1]

K. Kizilkaya,*† R. L. Fernando,* and D. J. Garrick*‡[2]

*Department of Animal Science, Iowa State University, Ames 50011; †Department of Animal Science, Adnan Menderes University, Aydin 09100 Turkey; and ‡Institute of Veterinary, Animal and Biomedical Science, Massey University, Palmerston North, New Zealand

ABSTRACT: Genomic prediction involves characterization of chromosome fragments in a training population to predict merit. Confidence in the predictions relies on their evaluation in a validation population. Many commercial animals are multibreed (MB) or crossbred, but seedstock populations tend to be purebred (PB). Training in MB allows selection of PB for crossbred performance. Training in PB to predict MB performance quantifies the potential for across-breed genomic prediction. Efficiency of genomic selection was evaluated for a trait with heritability 0.5 simulated using actual SNP genotypes. The PB population had 1,086 Angus animals, and the MB population had 924 individuals from 8 sire breeds. Phenotypic values were simulated for scenarios including 50, 100, 250, or 500 additive QTL randomly selected from 50K SNP panels. Panels containing various numbers of SNP, including or excluding the QTL, were used in the analysis. A Bayesian model averaging method was used to simultaneously estimate the effects of all markers on the panels in MB (or PB) training populations. Estimated effects were utilized to predict genomic merit of animals in PB (or MB) validation populations. Correlations between predicted and simulated genomic merit in the validation population was used to reflect predictive ability. Panels that included QTL were able to account for 50% or more of the within-breed genetic variance when the trait was influenced by 50 QTL. The predictive power eroded as the number of QTL increased. Panels that composed the QTL and no other markers were able to account for 50% or more genetic variance even with 500 QTL. Panels that included genomic markers as well as QTL had less predictive power as the number of markers on the panel was increased. Panels that excluded the QTL and relied on markers in linkage disequilibrium (LD) to predict QTL effects performed more poorly than marker panels with QTL. Real-life situations with 50K panels that excluded the QTL could account for no more than 20% genetic variation for 50 QTL and less than 10% for 500 QTL. The difference between panels that included and excluded QTL indicates that the predictive ability of existing panels is limited by their LD. Training in PB to predict MB tended to be more predictive than training in MB to predict PB due to greater average levels of LD in PB than in MB populations. Improved across breed prediction of genomic merit will require panels with more than 50,000 markers.

Key words: across-breed genomic prediction, Bayesian model averaging, marker density, single nucleotide polymorphism marker panel

J. Anim. Sci. 2010. 88:544–551
doi:10.2527/jas.2009-2064

INTRODUCTION

Genetic improvement of polygenic traits can be achieved through selection using breeding values estimated from phenotypes of the individual, its contemporaries, its relatives, or a combination. Selection is typically carried out in purebred (**PB**) populations to improve PB and crossbred performance under field conditions. Response to selection is proportional to the accuracy of predicted genetic merit (Falconer and Mackay, 1996). Recent developments enable dense genome-wide SNP to be used to increase prediction accuracies for young animals (Meuwissen et al., 2001) by exploiting linkage disequilibrium (**LD**) between markers and QTL. The prediction accuracy is positively related to LD.

[1]The authors acknowledge funding from USDA-CSREES-NRI 2008-56518-8726 (National Beef Cattle Evaluation Consortium) and 2009-35205-05100 (Bioinformatics to Implement Genomic Selection).

[2]Corresponding author: dorian@iastate.edu

Received April 21, 2009.
Accepted October 2, 2009.

Genomic selection (**GS**) of PB for crossbred performance involves estimating effects of SNP on crossbred performance, using phenotypes and genotypes from crossbreds, and applying the results to PB (Dekkers, 2007). Using simulated genotypic and phenotypic data, Toosi et al. (2008) and Ibánez-Escriche et al. (2009) found selection of PB for crossbred performance was successful by training in crossbred and multibreed (**MB**) populations even when breed-specific effects were ignored. In both studies the LD in training and target populations resulted from the manner in which ancestral populations were simulated and involved assumptions about effective population sizes, population bottlenecks, and mutation rates. The LD in simulations may not accurately reflect real beef cattle populations.

Actual high-density SNP genotypes exhibit LD that is not dependent upon assumed population sizes or mutation rates. This study uses actual SNP genotypes in PB and MB populations to simulate additive genetic merit and phenotypic performance for training and validation of GS. The value of PB training for MB performance and MB training for PB performance were compared.

MATERIALS AND METHODS

Animal Care and Use Committee approval was not obtained for this study because the data were obtained from an existing Iowa State University database of SNP genotypes.

SNP Data Sets

Illumina released a 50K SNP panel that included markers distributed across the bovine genome (Matukumalli et al., 2008). Genotypes from these panels were obtained from 2 resource populations. The first included 1,086 PB Angus animals from Iowa State University (Ames; Hassen et al., 2003). The second population was a subsample of animals from the Carcass Merit Project, a large MB study funded by the US beef cattle industry to evaluate the genetics of tenderness (Thallman et al., 2003). In that study, legacy AI sires from a range of breeds were mated to commercial cows, and DNA samples and phenotypes were collected from the steer offspring. The subsample of 924 animals that were genotyped represented 239 Angus-, 10 Brahman-, 183 Charolais-, 78 Hereford-, 45 Limousin-, 137 Maine-Anjou-, 97 Shorthorn-, and 135 South Devon-sired offspring. The Illumina A/B allele calls were used to compute a covariate for each locus that had values 0, 1, or 2 representing the number of B alleles. Missing genotypes represented less than 0.2% observations and were replaced with their breed average covariate. All genotypes were retained regardless of minor allele frequency.

Simulation of Additive Genetic Merit and Phenotypic Performance

A trait with heritability of 50%, determined by 50, 100, 250, or 500 additive bi-allelic QTL was simulated using Illumina 50K SNP genotypes from PB and MB populations as described below.

A random sample of $N = 50$, 100, 250, or 500 SNP were chosen from the observed 50K genotypes to represent QTL. Each locus had an equal probability of being included, regardless of minor allele frequency. Each QTL was assigned a parametric substitution effect by sampling from an independent normal distribution with mean 0 and variance $\frac{\sigma_g^2}{N2\overline{pq}}$ with $\overline{p}$ and $\overline{q}$ estimated from all SNP loci in the training population so that the expected additive genetic variance was σ_g^2 (Fernando et al., 2007). Residual effects for each animal were obtained by sampling from an independent normal distribution with mean 0 and variance $\sigma_e^2 = \sigma_g^2$ so that expected heritability was 0.5. Actual heritability would have varied between replicates and varied between the training and validation populations.

Additive genetic merit of each animal in the PB and MB populations was obtained as the sum of the substitution effects for each QTL allele. The simulated phenotypic performance of each animal was obtained by adding its residual effect to its additive genetic merit as follows:

$$y_i = \sum_{j=1}^{N} x_{ij}\alpha_j + e_i, \quad [1]$$

where y_i is the simulated phenotypic performance of animal i, $\sum_{j=1}^{N} x_{ij}\alpha_j$ is the additive genetic merit obtained by summing the genotypic values at each locus over all N loci, where the genotypic value at locus j is the product of the 0, 1, 2 covariate for animal i, x_{ij}, and the substitution α_j for that locus, and e_i is the random residual effect.

Five replicated data sets were generated for each of 4 QTL scenarios (QTL50, QTL100, QTL250, or QTL500) representing $N = 50$, 100, 250, or 500 QTL. In each replicate, a random set of SNP was chosen to be QTL, regardless of whether or not they had been sampled as QTL in other replicates or scenarios. The substitution effects were sampled separately in each replicate but applied equally to the training and validation populations for any particular replicate.

The SNP locus with the highest LD (**HLD**), regardless of genomic location, was identified for each QTL from among the 50K SNP markers excluding those chosen to be QTL. The LD between 2 loci was quantified in this study as the square of the correlation between the 2 covariates.

Definition and Use of Marker Panels

Five marker panels comprising observed SNP genotypes were defined for the analysis of each replicate of the simulated phenotypes for each QTL scenario: panel 1 (QTL), portfolio of QTL; panel 2 (QTL and HLD), the portfolio of QTL and corresponding HLD loci; panel 3 (50K and QTL), all 50K markers, including the QTL; panel 4 (HLD), only the HLD markers; panel 5 (50K without QTL), all 50K markers, excluding the QTL. For example, in the QTL50 scenario, there were 50 markers in panel 1, 100 in panel 2, 50K in panel 3, 50 in panel 4, or 50K-50 in panel 5.

Each analysis involved estimation of SNP effects from 1 source of data (PB or MB), known as training, that were subsequently used to predict genetic merit in the other source, known as validation. The accuracy of prediction was quantified in training and validation using the correlation between the predicted and simulated additive genetic merit. In addition, the correlation between predicted additive genetic merit and simulated phenotype was calculated in the training data to measure the extent covariates were predicting residual effects rather than additive genetic merit. Correlations were averaged across the 5 replicates for any particular combination of panels and scenarios.

Bayesian Estimation of SNP Effects

Marker effects were estimated in the training population using the method described here that is similar to so-called Bayes-B (Meuwissen et al., 2001). It is based on the model

$$\mathbf{y} = 1\mu + \sum_{j=1}^{K} \mathbf{x}_j \beta_j \delta_j + \mathbf{e}, \qquad [2]$$

where $\mathbf{y}$ is the vector of simulated phenotypic values, μ is an overall mean, K is the number of marker loci in the panel, $\mathbf{x}_j$ is the column vector representing the covariate at locus j, β_j is the random substitution effect for locus j, which is conditional on σ^2_β and is assumed normally distributed $N(0, \sigma^2_\beta)$ when $\delta_j = 1$ but $\beta_j = 0$ when $\delta_j = 0$, δ_j is a random 0/1 variable indicating the absence (with probability π) or presence (with probability $1 - \pi$) of locus j in the model, and $\mathbf{e}$ is the vector of random residual effects assumed normally distributed $N(0, \sigma^2_e)$. The variance σ^2_β (or σ^2_e) was a priori assumed to be scaled inverse chi-squared with $\nu_\beta = 4$ (or $\nu_e = 10$) df and scale parameter S^2_β (or S^2_e). The known values for parameter π were 0 for panel 1 (QTL) and panel 4 (HLD), 0.5 for panel 2 (QTL and HLD), and 0.999, 0.998, 0.995, or 0.99 for QTL50, QTL100, QTL250, or QTL500 in panels 3 (50K and QTL) and 5 (50K without QTL), respectively, representing 1 minus the ratio of QTL to markers.

The expected value of a scaled inverse chi-squared random variable σ^2 with scale parameter S^2 and degrees of freedom ν is $E(\sigma^2) = \frac{S^2\nu}{\nu - 2}$, so the scale parameter $S^2 = \frac{E(\sigma^2)(\nu - 2)}{\nu}$. Recall that the substitution effects were sampled from a normal distribution with mean 0 and variance $\sigma^2_\beta = \frac{\sigma^2_g}{N2\overline{pq}}$, where N can be written as $K(1-\pi)$. Applying these facts leads to the scale parameter for the variance of the substitution effects to be $S^2_\beta = \frac{\sigma^2_g(\nu - 2)}{K(1-\pi)2\overline{pq}\nu}$, which gives $E(\sigma^2_\beta) = \frac{\sigma^2_g}{K(1-\pi)2\overline{pq}}$, and similarly the scale parameter for the residual variance to be $S^2_e = \frac{\sigma^2_e(\nu - 2)}{\nu}$. The parametric values for σ^2_g and σ^2_e required to derive the scale factors were those used to simulate the data. However, these variables were treated as unknowns in the analyses so that the posterior means for marker substitution effects were obtained by marginalizing over the unknown genetic and residual variances; note that σ^2_β is common to all loci in the model in contrast to the locus specific variance components in Bayes-B (Meuwissen et al., 2001).

Marker effects β_j were estimated by computing Monte-Carlo means of the posterior distribution of these effects, as described below, using a Gibbs sampling strategy. At a locus j, samples for δ_j were drawn from its conditional distribution given μ, the effects at all other loci included in the model, and the 2 variance components. This conditional distribution can be written as

$$\begin{aligned}\Pr(\delta_j \mid \mathbf{y}, \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e) &= \frac{f(\delta_j, \mathbf{y} \mid \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e)}{f(\mathbf{y} \mid \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e)} \\ &= \frac{\Pr(\delta_j) f(\mathbf{y} \mid \delta_j, \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e)}{f(\mathbf{y} \mid \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e)},\end{aligned} \qquad [3]$$

where $\boldsymbol{\beta}_{-j}$ is the vector of effects in the model other than at locus j. The denominator of [3] is obtained as the sum of the numerator computed for $\delta_j = 0$ and $\delta_j = 1$. When $\delta_j = 0$, $\Pr(\delta_j)$ is equal to π and $f(\mathbf{y} \mid \delta_j, \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e)$ is a normal density with mean

$$E(\mathbf{y} \mid \delta_j = 0, \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e) = 1\mu + \sum_{j' \neq j} \mathbf{x}_{j'} \beta_{j'} \delta_{j'} \qquad [4]$$

and variance

$$\mathrm{Var}(\mathbf{y} \mid \delta_j = 0, \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e) = \mathbf{I}\sigma^2_e. \qquad [5]$$

Table 1. Correlations between estimated genotypic values ($\hat{g}_{Training\ \ data}$) and phenotypes ($y$) and true genotypic values (g) in multibreed (MB) and purebred (PB) training data sets[1]

QTL scenario[2] and marker panel[3]	$r(y_{MB}, \hat{g}_{MB})$	$r(g_{MB}, \hat{g}_{MB})$	$r(y_{PB}, \hat{g}_{PB})$	$r(g_{PB}, \hat{g}_{PB})$
QTL50				
QTL	0.701	0.965	0.676	0.978
QTL and HLD	0.713	0.949	0.683	0.966
50K and QTL	0.791	0.848	0.744	0.903
HLD	0.500	0.641	0.516	0.719
50K without QTL	0.783	0.706	0.736	0.755
QTL100				
QTL	0.733	0.944	0.710	0.956
QTL and HLD	0.747	0.930	0.724	0.937
50K and QTL	0.871	0.790	0.826	0.827
HLD	0.533	0.643	0.567	0.734
50K without QTL	0.855	0.715	0.815	0.747
QTL250				
QTL	0.772	0.883	0.747	0.905
QTL and HLD	0.795	0.855	0.759	0.883
50K and QTL	0.926	0.736	0.882	0.776
HLD	0.646	0.685	0.640	0.753
50K without QTL	0.906	0.712	0.872	0.756
QTL500				
QTL	0.813	0.810	0.813	0.877
QTL and HLD	0.835	0.777	0.826	0.858
50K and QTL	0.949	0.698	0.897	0.781
HLD	0.671	0.612	0.700	0.762
50K without QTL	0.942	0.691	0.885	0.772

[1]Correlations are average of 5 replicates for $h^2 = 0.5$.

[2]QTL50, QTL100, QTL250, and QTL500 represent simulated phenotypes based on 50, 100, 250, and 500 QTL.

[3]QTL: only QTL genotypes in the model; QTL and HLD: QTL and markers with highest linkage disequilibrium (LD) to QTL in the model; 50K and QTL: QTL and 50K markers in the model; HLD: only markers with highest LD to QTL in the model; and 50K without QTL: only 50K markers in the model.

When $\delta_j = 1$, $\Pr(\delta_j)$ is equal to $1 - \pi$ and $f(\mathbf{y} \mid \delta_j, \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e)$ is a normal distribution with the same mean as with $\delta_j = 0$, but with variance

$$\text{Var}(\mathbf{y} \mid \delta_j = 1, \boldsymbol{\beta}_{-j}, \sigma^2_\beta, \sigma^2_e) = \mathbf{x}_j \mathbf{x}'_j \sigma^2_\beta + \mathbf{I}\sigma^2_e. \quad [6]$$

These normal densities can be computed efficiently using mixed model methods (Henderson, 1984). If the sampled value for $\delta_j = 1$, then the value of β_j was sampled from its full conditional, which is normal with mean $\hat{\beta}_j$ and variance $\frac{\sigma^2_e}{c}$ (Sorensen and Gianola, 2002), where $c = (\mathbf{x}'_j \mathbf{x}_j + \frac{\sigma^2_e}{\sigma^2_\beta})$, and $\hat{\beta}_j = \frac{\mathbf{x}'_j(\mathbf{y} - 1\mu - \sum_{j' \neq j} \mathbf{x}_{j'} \beta_{j'} \delta_{j'})}{c}$. Samples for μ, σ^2_β, and σ^2_e were obtained from their full conditional distributions (Sorensen and Gianola, 2002). This procedure was implemented in GenSel (Fernando and Garrick, 2008).

For each replicated data set within each QTL scenario, a burn-in period of 5,000 Markov chain Monte Carlo (**MCMC**) cycles was used before saving samples from each of an additional 40,000 MCMC cycles. The mean substitution effect of each marker obtained from the post-burn samples were multiplied by its covariate values for each animal, and summed across all markers in the panel to estimate genomic merit. This was done for animals in the training and validation populations.

RESULTS AND DISCUSSION

Training Results

Table 1 shows some statistics related to the performance of genomic predictions when applied in the same population used for training. In real life, we will not know in advance how many QTL contribute to a particular trait. The number of QTL (QTL scenario) will be dictated by the biology of the trait. The researcher, however, will have the opportunity to determine the nature of the marker panel used in genotyping a population. The correlations between estimated and true genetic merit (columns 3 and 5) are the critical parameters from the viewpoint of selection. Table 1 in the QTL50 scenario shows that genomic prediction can achieve high correlations when the marker panel includes QTL (QTL, QTL+HLD, QTL+50K). Panels that include redundant markers do a poorer job of predicting merit than panels that contain only QTL. The more redundant markers, the poorer the prediction, al-

though the reduction in correlation is only small when only the HLD markers are added to the panel. Including all 50K markers results in type I errors that reduce the predictive ability of the panel. That is, spurious markers are included in the model and therefore contribute to the prediction of merit, even though the analysis did not involve hypothesis testing. The probability that variable $\delta_j = 1$ determines the impact of each locus could in theory approach 0 for loci that are completely uninformative. However, in practice the frequency of that variable is seldom exactly equal to 0 (results not shown), and 50K loci with small spurious effects can collectively introduce prediction errors. In the worst case, panel QTL+50K, the prediction still accounts for 72% ($r^2 = 0.848^2$) or more of genetic variation.

In real life, current marker panels are unlikely to include causal mutations. Accordingly, panels HLD and 50K without QTL are more realistic, representing reduced panels constructed from previously discovered QTL regions and high-density genomic panels, respectively. These panels that omit QTL perform substantially poorer than the panels that include QTL. The panel that included every SNP (50K without QTL) performed slightly better than the panel that included only the HLD markers. Panels with more SNP have a greater chance of including markers in LD with QTL and allow the possibility of predicting each QTL from the collective action of several markers.

Comparing the QTL and HLD panels, the results demonstrate a moderate reduction in correlation (e.g., in MB 0.96 vs. 0.64) that corresponds to a considerable reduction in genetic variance accounted for by markers (e.g., from 92 to 41%). This reduction would not have been as large if the marker panel included loci with greater LD.

Table 1 shows the correlations of genomic prediction with phenotypic performance and those correlations tend to increase with the number of markers on the panel. Comparing these correlations to those between estimated and true genomic merit demonstrates that nongenetic or residual effects are being predicted in these large panels. The fact that these nongenetic effects are being attributed to loci is associated with the type I errors described previously.

The comparison of the 4 scenarios in Table 1 shows that genetic correlations between true and estimated merit tend to decline as the number of QTL increase. The simulation held the genetic variation constant across scenarios, so an increase in the number of QTL is associated with a decrease in the average size of the QTL.

Generally speaking, training in PB resulted in slightly greater correlations of predicted and actual genetic merit than did training in MB. In contrast, training in MB resulted in slightly greater correlations between predicted genetic merit and phenotype, indicating that the MB analysis was predicting nongenetic or residual effects to a greater extent than PB. There is no obvious explanation for this phenomenon.

Table 2. Correlations between true (g) and predicted ($\hat{g}_{Validation\ data|Training\ data}$) genotypic values in the validation data set (MB and PB) after estimating the substitution effect in an independent training data set[1,2]

QTL scenario[3] and marker panel[4]	$r(g_{PB}, \hat{g}_{PB\|MB})$	$r(g_{MB}, \hat{g}_{MB\|PB})$
QTL50		
QTL	0.953	0.962
QTL and HLD	0.931	0.938
50K and QTL	0.766	0.842
HLD	0.570	0.486
50K without QTL	0.388	0.422
QTL100		
QTL	0.938	0.941
QTL and HLD	0.914	0.898
50K and QTL	0.585	0.665
HLD	0.513	0.480
50K without QTL	0.289	0.308
QTL250		
QTL	0.840	0.853
QTL and HLD	0.788	0.785
50K and QTL	0.399	0.425
HLD	0.510	0.429
50K without QTL	0.247	0.276
QTL500		
QTL	0.720	0.786
QTL and HLD	0.642	0.710
50K and QTL	0.254	0.384
HLD	0.372	0.391
50K without QTL	0.200	0.299

[1]Correlations are average of 5 replicates for $h^2 = 0.5$.
[2]MB = multibreed population; PB = purebred population.
[3]QTL50, QTL100, QTL250, and QTL500 represent simulated phenotypes based on 50, 100, 250, and 500 QTL.
[4]QTL: only QTL genotypes in the model; QTL and HLD: QTL and markers with highest linkage disequilibrium (LD) to QTL in the model; 50K and QTL: QTL and 50K markers in the model; HLD: only markers with highest LD to QTL in the model; and 50K without QTL: only 50K markers in the model.

In practice, particular interest is in the reliability of predicting animals that are not in the training population. This is because training populations are used to predict the merit of new selection candidates at young ages when they will have at most only prepubertal phenotypes and will not have had the opportunity to produce any offspring. Table 2 shows the results of cross-validation, training in MB to predict PB, and training in PB to predict MB. Training in MB to predict PB is of interest because it follows the theoretical simulations undertaken by Ibánez-Escriche et al. (2009) and Toosi et al. (2009) for evaluating the efficacy of genomic prediction for the purpose of selecting PB animals for commercial crossbred performance (Dekkers, 2007). Training in PB to predict MB performance is of interest because the genotyping and phenotyping associated with training is an expensive process and there is potential to use training analyses from 1 breed to predict the performance of other breeds.

Table 3. Correlations[1] between true and predicted genotypic values from training in purebred Angus (PB) and validating in individual sire breeds that composed the multibreed (MB) population

QTL scenario[2] and marker panel[3]	Angus	Brahman	Charolais	Hereford	Limousin	Maine-Anjou	Shorthorn	Southdevon
QTL50								
QTL	0.972	0.969	0.956	0.946	0.955	0.967	0.962	0.958
QTL and HLD	0.955	0.919	0.936	0.894	0.908	0.942	0.944	0.929
50K and QTL	0.859	0.858	0.849	0.811	0.810	0.854	0.820	0.836
HLD	0.608	0.339	0.425	0.408	0.412	0.488	0.421	0.481
50K without QTL	0.511	0.227	0.382	0.336	0.355	0.390	0.437	0.403
QTL100								
QTL	0.948	0.918	0.951	0.918	0.938	0.936	0.946	0.930
QTL and HLD	0.919	0.870	0.905	0.847	0.915	0.903	0.904	0.871
50K and QTL	0.704	0.582	0.668	0.548	0.659	0.663	0.657	0.665
HLD	0.623	0.378	0.425	0.426	0.545	0.441	0.422	0.377
50K without QTL	0.386	0.226	0.284	0.158	0.366	0.244	0.363	0.254
QTL250								
QTL	0.882	0.881	0.867	0.816	0.834	0.860	0.825	0.806
QTL and HLD	0.835	0.665	0.781	0.743	0.746	0.805	0.760	0.718
50K and QTL	0.479	0.144	0.415	0.463	0.427	0.415	0.474	0.289
HLD	0.590	−0.197	0.401	0.427	0.242	0.414	0.337	0.310
50K without QTL	0.404	−0.026	0.237	0.300	0.193	0.196	0.306	0.104
QTL500								
QTL	0.819	0.754	0.783	0.739	0.831	0.771	0.808	0.730
QTL and HLD	0.760	0.530	0.704	0.680	0.764	0.705	0.694	0.628
50K and QTL	0.496	0.239	0.380	0.197	0.400	0.260	0.385	0.327
HLD	0.530	0.084	0.331	0.301	0.400	0.365	0.300	0.313
50K without QTL	0.435	0.124	0.262	0.118	0.276	0.174	0.263	0.247

[1]Correlations are average of 5 replicates for $h^2 = 0.5$.

[2]QTL50, QTL100, QTL250, and QTL500 represent simulated phenotypes based on 50, 100, 250, and 500 QTL.

[3]QTL: only QTL genotypes in the model; QTL and HLD: QTL and markers with highest linkage disequilibrium (LD) to QTL in the model; 50K and QTL: QTL and 50K markers in the model; HLD: only markers with highest LD to QTL in the model; and 50K without QTL: only 50K markers in the model.

Validation Results

The trends observed in Table 1 are typically retained in Table 2. That is, panels that include QTL do better than panels that rely entirely on LD. However, in cross-validation (Table 2), the panels that exclude the QTL do not perform as well as in training (Table 1). The scenarios that involved a larger number of QTL were characterized by greater erosion of predictive performance in cross-validation than in training. Both of these observations are attributed to differences in LD between the training and validation populations.

Table 2 demonstrates that the proportion of genetic variation that can be accounted using training populations of around 1,000 individuals with the currently available 50K marker panel applied across breed, declines from 15 to 18% when 50 QTL are responsible for variation in a trait, to only 4 to 9% when there are 500 underlying QTL. It could be argued that the minor allele frequencies of real QTL are likely to be less than those of SNP on the marker panel. Because the maximum possible LD between a pair of loci is limited by the difference in their allele frequency, the simulated results will represent optimistic results.

The MB population represented offspring sired by 8 breeds, including Angus. The correlation between training in PB and simulated merit of each individual breed is shown in Table 3. These results by sire breed demonstrate greater correlations in the Angus than in the other breeds for all scenarios and marker panels. The scenario involving 500 QTL with the most realistic marker panel, 50K without QTL, achieved a correlation of 0.435 in Angus, accounting for less than 20% of genetic variation. Correlations in the other breeds ranged from 0.118 to 0.276 and accounted for less than one-third of variation achieved in Angus. Individual replicates exhibited variation in response (results not shown) such that chance sampling led to poorer performance in some breeds for the scenarios involving 100 or 250 QTL compared with 500 QTL. The Brahman breed

Table 4. Average linkage disequilibrium (LD) between loci selected to represent QTL and the most informative marker from the 50K panel in purebred (PB) or multibreed (MB) populations

	Marker-QTL LD assessed in	
Item	PB	MB
Marker highest LD to QTL chosen from		
PB	0.549[1]	0.322
MB	0.412	0.408

[1]Diagonals represent LD in training populations and off-diagonals represent LD in validation populations (depicted by the column) for markers chosen in training populations (depicted by the row).

was only represented by 10 offspring and resulted in negative correlations between predicted and simulated merit for 2 of the marker panels. The *Bos indicus* breed is genetically more distant from Angus than the other *Bos taurus* breeds and is accordingly expected to have more disparate pairwise LD than Angus.

The simulation considered additive effects, ignoring dominance and epistasis. Inclusion of those modes of gene action would be expected to erode the predictive performance across breeds that vary in gene frequency.

The relative performance of training in PB in relation to MB demonstrated that PB was a better choice, given the size of the training sets and the current marker density. Again, we attribute this fact to differences in LD between PB and MB populations.

The diagonal elements in Table 4 quantify the LD between the QTL and HLD markers in the PB and MB populations. The HLD markers were more highly predictive of the QTL in PB than in MB. Nevertheless, the average LD for all QTL for the highest markers on the panel in PB was only 0.549. Inbred populations are, by definition, characterized by smaller effective population sizes than outbred populations. Accordingly, the LD is greater in PB than in the more outbred MB population. The average LD of the HLD for the MB was 0.408, considerably less than PB. The off-diagonals in Table 4 characterize the extent to which LD in 1 population was consistent with LD in the other population. The average LD observed between QTL and HLD markers in the PB training population (0.549) was reduced when compared with the average value for the same pairs of loci in the MB population (0.322). In contrast, the reverse was true when HLD markers were chosen from MB. The average LD for the HLD markers increased from 0.408 in the MB population to 0.412 in the PB.

Figure 1 further depicts, in terms of the correlation between HLD marker and QTL, the pairwise relationship between HLD panels chosen in a training population (x-axis) and applied in a validation population (y-axis) based on 1 replicate from the QTL500 scenario. Ideal marker panels would have sufficient LD that any QTL would have a high LD counterpart on the panel. Such markers might have a positive or a negative correlation given the arbitrary nature of their A/B allele labeling. Accordingly, the figure depicts the absolute value of the correlation in the training population on the x-axis. The original sign of the correlation in the validation population has been retained, so that markers with opposite phase appear with negative correlations on the y-axis. Ideal markers would cluster with x-coordinates close to 1. The figure shows that many of the HLD markers have correlations less than 0.5, some less than 0.4, more so when the markers are chosen in the MB population. Ideally, the LD in one population would be retained in the other, such that the points would converge on the line y = x, which is overlain on the graph. The left graph shows that the points tend to fall below the line y = x, demonstrating for markers chosen in the PB, reduced LD in the MB than in the PB. In contrast, the right graph depicts points closer to the line y = x, demonstrating that LD in MB tends to

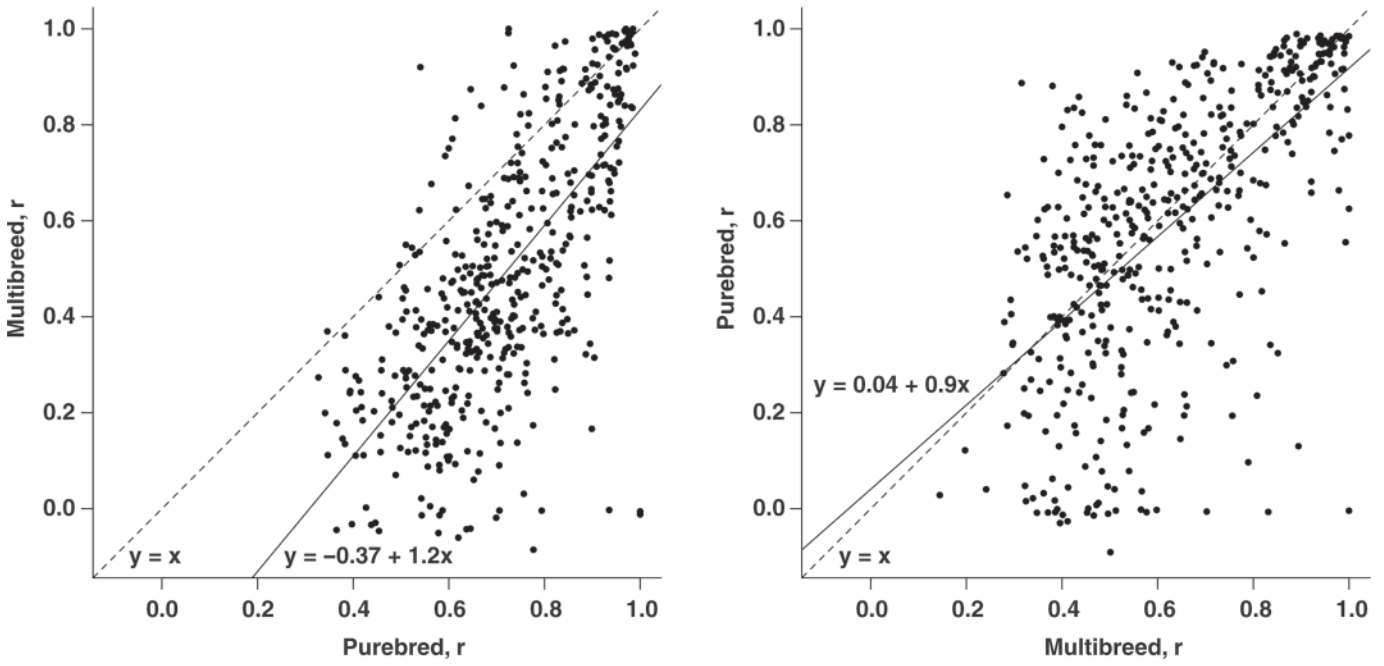

Figure 1. Plots of correlations (r) between QTL and highest linkage disequilibrium (HLD) markers from purebred (left figure) and multibreed (right figure) training populations on the x-axis against correlation values for the same pair of QTL-HLD markers for multibreed (and purebred) populations on the y-axis. The sign of the correlation in the training population is arbitrary according to the allele coding, so all correlations in the training population (x-axis) are depicted as positive.

be consistent with LD in PB, but closer to the origin, demonstrating the reduced average levels of LD in MB compared with PB.

In PB populations, increased LD is retained over greater genomic distances than in MB. Accordingly, HLD markers chosen in PB may be some distance from the QTL they are marking. Such markers will be less informative when translated to another breed. In MB populations, LD is retained more locally than PB. In that case, HLD markers chosen in a MB population will be close to the QTL and therefore still likely to work in PB populations; in fact, they may even be greater in the PB than the MB. In summary, it is harder to find HLD markers in MB populations because there is overall less LD, but high LD that is observed is likely on average to be even higher in PB populations.

Collectively, these findings suggest that denser SNP panels with greater LD than is available on the Illumina 50K panel will be an advantage for genomic training, especially in selecting PB for crossbred performance. Haplotype strategies may provide an in-silico method of increasing LD and need further research in the context of within- and across-breed genomic prediction.

LITERATURE CITED

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

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

Fernando, R. L., and D. J. Garrick. 2008. GenSel—User manual for a portfolio of genomic selection related analyses. Animal Breeding and Genetics, Iowa State University, Ames. http://taurus.ansci.iastate.edu/gensel Accessed Apr. 21, 2009.

Fernando, R. L., D. Habier, C. Sticker, J. C. M. Dekkers, and L. R. Totir. 2007. Genomic selection. Acta Agriculturae Scand. Section A 57:192–195.

Hassen, A., D. E. Wilson, and G. H. Rouse. 2003. Estimation of genetic parameters for ultrasound-predicted percentage of intramuscular fat in Angus cattle using random regression models. J. Anim. Sci. 81:35–45.

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

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

Matukumalli, L. K., R. D. Schnabel, C. T. Lawley, T. S. Sonstegard, T. P. L. Smith, S. S. Moore, J. F. Taylor, and C. P. Van Tassell. 2008. Characterization of the cattle HapMap population using the Illumina Bovine-50K SNP chip. Proc. Plant and Animal Genome XVI. San Diego, CA.

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

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

Thallman, R. M., D. W. Moser, E. W. Dressler, L. R. Totit, R. L. Fernando, S. D. Kachman, J. M. Rumph, M. E. Dikeman, and E. J. Pollak. 2003. Carcass merit project: DNA marker validation. Proc. Beef Improv. Fed. 8th Genet. Prediction Workshop 8:70–90.

Toosi, A., R. L. Fernando, J. C. M. Dekkers, and R. L. Quaas. 2008. Genomic selection for purebreds using data from admixed populations. J. Anim. Sci. 86(E-Suppl. 2):361. (Abstr.)

Toosi, A., R. L. Fernando, J. C. M. Dekkers, and R. L. Quaas. 2009. Genomic selection in admixed and crossbred populations. J. Anim. Sci. 88:32–46.