Accuracy of genomic prediction using low-density marker panels

Citation for published version:
Zhang, Z, Ding, X, Liu, J, Zhang, Q & de Koning, DJ 2011, 'Accuracy of genomic prediction using low-density marker panels' Journal of Dairy Science, vol. 94, no. 7, pp. 3642-3650. DOI: 10.3168/jds.2010-3917

Digital Object Identifier (DOI):
10.3168/jds.2010-3917

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Journal of Dairy Science

J. Dairy Sci. 94:3642–3650
doi:10.3168/jds.2010-3917

Accuracy of genomic prediction using low-density marker panels

Z. Zhang,*† X. Ding,* J. Liu,* Q. Zhang,*[1] and D.-J. de Koning†‡[1]
*Key Laboratory of Animal Genetics and Breeding of the Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing, 100193, China
†The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, EH25 9RG, UK
‡Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, 750 07 Uppsala, Sweden

ABSTRACT

Genomic selection has been widely implemented in national and international genetic evaluation in the dairy cattle industry, because of its potential advantages over traditional selection methods and the availability of commercial high-density (HD) single nucleotide polymorphism (SNP) panels. However, this method may not be cost-effective for cow selection and for other livestock species, because the cost of HD SNP panels is still relatively high. One possible solution that can enable other species to benefit from this promising method is genomic selection with low-density (LD) SNP panels. In this simulation study, LD SNP panels designed with different strategies and different SNP densities were compared. The effects of number of quantitative trait loci, heritability, and effective population size were evaluated in the framework of genomic selection with LD SNP panels. Methodologies of Bayesian variable selection; BLUP with a trait-specific, marker-derived relationship matrix; and BLUP with a realized relationship matrix were employed to predict genomic estimated breeding values with both HD and LD SNP panels. Up to 95% of accuracy obtained by using an HD panel can be obtained by using only a small proportion of markers. The LD panel with markers selected on the basis of their effects always performs better than the LD panel with evenly spaced markers. Both the genetic architecture of the trait and the effective population size have a significant effect on the performance of the LD panels. We concluded that, to implement genomic selection with LD panels, a training population of sufficient size and genotyped with an HD panel is necessary. The trade-off between the LD panels with evenly spaced markers and selected markers must be considered, which depends on the number of target traits in a breeding program and the genetic architecture of these traits. Genomic selection with LD panels could be feasible and cost-effective, though before implementation, a further detailed genetic and economic analysis is recommended.

Key words: low-density panel, genomic selection, TABLUP, BayesB

INTRODUCTION

The recent development of high-throughput genotyping technology for cattle and other species has facilitated the implementation of genomic selection (**GS**; Meuwissen et al., 2001). Because of its potential advantages, GS has been widely implemented in national and international dairy cattle genetic evaluation (Harris and Montgomerie, 2009; Hayes et al., 2009; VanRaden et al., 2009). The complete implementation of GS in dairy bull selection could be financially attractive for breeding organizations (Schaeffer, 2006; König et al., 2009). However, it requires high-density (**HD**) SNP panels genotyped on a large training population and on all candidates whose genetic merit needs to be evaluated. Currently, the cost of the HD panel genotyping for such a large population is still very high. This high cost potentially prevents the broader implementation of GS in cows and other livestock species such as chicken or sheep, in which the individual selection candidates are not as economically important as dairy bulls (Goddard and Hayes, 2009).

One solution to enable cows and other livestock species to benefit more from genomic selection is predicting the genomic EBV (**GEBV**) with low-density (**LD**) panels. Habier et al. (2009) presented a method that can use the co-segregation information from LD panels with evenly spaced or selected markers to track marker effects of HD panels within families. This method can effectively preserve the accuracy of genomic prediction with LD panels for individuals with both parents being genotyped with HD panels (Habier et al., 2009; VanRaden et al., 2010). However, in practice, genotyping the parents of all selection candidates with HD panels might be still too expensive, especially for species with low reproduction rates. If not all parents of candidates could be genotyped with HD panels, the loss of accuracy with LD panels would be potentially

Received October 10, 2010.
Accepted March 15, 2011.

[1] Corresponding authors: qzhang@cau.edu.cn and dj.de-koning@hgen.slu.se

high (Habier et al., 2009; VanRaden et al., 2010). All methods that have been proposed in the framework of genomic selection with HD panels can also be directly applied to LD panels. Cleveland et al. (2010) compared 3 Bayesian methods using subsets of simulated markers from the common data set of the thirteenth QTL-MAS Workshop (Coster et al., 2010) and found the Bayesian variable selection method with student t-distribution gave the best estimate with a subset of markers selected from that data set. In dairy cattle, Weigel et al. (2009) compared the predicting ability of different LD panels in a large data set. The LD panels with selected markers performed better than the evenly spaced LD panels for the lifetime net merit. Additionally, researchers have worked with LD marker panels (Long et al., 2007; González-Recio et al., 2008; González-Recio et al., 2009; Vazquez et al., 2010). However, several factors affect the performance of genomic prediction with LD panels. The relative advantage between LD panels with evenly spaced and selected markers for different methods and trait architectures is yet to be investigated.

In the present study, the LD with evenly spaced (**ELD**) or selected (**SLD**) markers were designed with HD panels as the training panel. The effect of number of QTL ($\boldsymbol{N}_{\mathbf{QTL}}$), heritability, and effective population size were investigated. The performance of Bayesian variable selection method B (**BayesB**; Meuwissen et al., 2001), BLUP with a realized relationship matrix (**GBLUP**; VanRaden, 2008) and BLUP with a trait-specific, marker-derived relationship matrix (**TABLUP**; Zhang et al., 2010) were compared. The strategies to implement genomic selection with LD marker panels were evaluated within a wide range of population parameters.

MATERIALS AND METHODS

Data Simulation

The simulation started with an initial population of N_e individuals and followed by $4N_e$ discrete generations, denoted as historical generations. In the initial population and each historical generation, males and females were randomly selected to form N_e matings and produced N_e offspring with $0.5N_e$ males and $0.5N_e$ females. This gave an effective population size of N_e. After the historical generations, 2 additional generations were simulated and denoted as training population and validation population, respectively. In the training population, the population size was expanded from N_e to 1,000 by random mating. Then, the 500 males were randomly mated with the 500 females, with each mating producing 2 offspring (1 male and 1 female) to obtain the validation population.

The simulated genome consisted of 10 chromosomes with a total length of 10 morgans (1 morgan per chromosome). On each chromosome, 1,000 marker loci were randomly located and each segment between 2 adjacent markers was considered a potential QTL, giving 10,000 marker loci and 9,990 potential QTL in total. Between 2 adjacent loci on the same chromosome, Haldane's mapping function was used to calculate the probability of recombination.

The mutation-drift equilibrium model was used to simulate polymorphic markers and QTL. Under this model, the expected heterozygosity is $H_e = 4N_eu/(1 + 4N_eu)$ when the population reaches its mutation-drift equilibrium (Jacquard, 1974). With different N_e, the mutation rates (u) were selected to give an H_e of 0.5 to ensure a high proportion of polymorphic markers in the historical generations. For example, with $N_e = 1{,}000$, the mutation rate would be 2.5×10^{-4} per locus, per generation, and per animal. Mutation was allowed throughout the historical generations for all loci. In the initial population, all markers and QTL had both alleles coded as 1. For each new mutation on the same locus, a unique allele was created and coded with a new number starting from 2. In the last historical generation, recoding of alleles was implemented to obtain bi-allelic markers. For each locus, the allele with the highest frequency was recoded as 1, whereas all other alleles were recoded as 2. This recoding strategy is similar to that of Solberg et al. (2008), with the difference that in Solberg et al. (2008), among all mutated alleles, the one with the highest frequency was treated as the visible mutated allele, whereas all others were treated as invisible mutations and, thus, had the same code as the ancestral allele. No mutations happened in the training and validation populations. The linkage disequilibrium (r^2; Hill and Robertson, 1968) between any pair of markers in the training population was calculated for the case of $N_e = 100$ and plotted against the marker distance in Figure 1. The average r^2 between 2 adjacent markers was 0.32.

True breeding values (**TBV**) were generated for all individuals in the training and validation populations. For each individual, TBV was obtained by summing up the effects of all QTL $\left(\text{i.e., } TBV = \sum_{j=1}^{m} z_j a_j\right)$, where a_j is the effect of QTL j, which was drawn from a gamma distribution with the shape parameter $\beta = 0.4$ and scale parameter $\alpha = 1.66$ [following Meuwissen et al. (2001)], m is the total number of QTL, and z_j equals -1, 0, or 1 for genotype 11, 12, and 22, respectively. Different numbers of QTL were randomly selected from the 9,990 putative QTL. Following Daetwyler et al. (2010), the N_{QTL} can be expressed relative to the number of effective chromosome segments ($\boldsymbol{M_e}$), which is

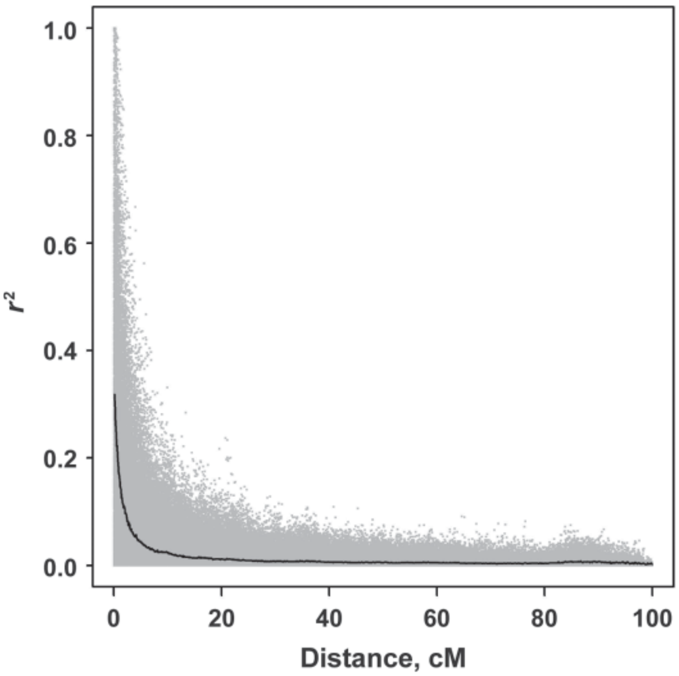

Figure 1. Linkage disequilibrium (r^2) between marker pairs in relation to the marker distances when effective population size (N_e) = 100. The solid line indicates the mean r^2 over successive intervals.

estimated by $M_e = 2N_eL/\ln(4N_eL)$ (Goddard, 2009), where L is the length of genome in morgan. Three levels of N_{QTL} relative to M_e were considered (i.e., low: $N_{\text{QTL}} = 0.05M_e$; medium: $N_{\text{QTL}} = 0.3M_e$; and high: $N_{\text{QTL}} = 1M_e$; Table 1).

Only the 1,000 individuals in the training population were assigned a phenotypic record. The phenotypic values of the ith individual, p_i, were obtained by $p_i = TBV_i + e_i$, where e_i is randomly sampled from the normal distribution with $N\left(0, \sigma_e^2\right)$. The total genetic variance was computed as the sum of variances across all QTL with the assumption of no correlation between QTL. The simulated additive genetic variance of each QTL was calculated as $\sigma_{g_j}^2 = 2p_j(1 - p_j)a_j^2$ (Falconer and Mackay, 1996), where p_j is the allele frequency at QTL j in the training population. The allele substitution effects were re-scaled to have an overall additive genetic variance $\left(\sigma_A^2\right)$ of 1. The environmental variance $\left(\sigma_e^2\right)$ was generated from $\left(1 - h^2\right)\sigma_A^2 \Big/ h^2$ to get the desired heritability. In the standard scenario, the heritability was 0.5.

Table 1. Number of simulated QTL in relation to the number of effective chromosome segments

		N_{QTL}[3]		
N_e[1]	M_e[2]	Low	Medium	High
50	131.6	6	39	131
100	241.1	12	72	241
200	445.1	22	133	445
500	1,009.7	50	302	1,009
1,000	1,887.4	94	566	1,887

[1] Effective population size.

[2] Number of effective chromosome segments, estimated by $M_e = 2N_eL/\ln(4N_eL)$, where L is the length of the genome in morgans, which is 10 in this study.

[3] Number of simulated QTL; low: $N_{\text{QTL}} = 0.05M_e$, medium: $N_{\text{QTL}} = 0.3M_e$, and high: $N_{\text{QTL}} = 1M_e$.

In the standard scenario in the simulation, N_e was set to be 100 and h^2 was 0.5. To investigate the effect of effective population size and heritability on the accuracies of genomic selection with LD panels, 2 groups of alternative scenarios were simulated in addition to the standard scenario. In the first group, 4 heritabilities were simulated: 0.1, 0.3, 0.8, and 0.95. In the second group, 4 effective population sizes were simulated: 50, 200, 500, and 1,000. For all of these alternatives, only the intended parameter was altered from the standard scenario. For all scenarios, 10 replicates were simulated.

Estimation of Marker Effects

The BayesB method (Meuwissen et al., 2001) was used to estimate marker effects in the training population. The statistical model can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{i=1}^{N} \mathbf{z}_i g_i + \mathbf{e}, \qquad [1]$$

where $\mathbf{y}$ is the vector of phenotypic values, $\mathbf{b}$ is a vector of fixed effects (including an overall mean), g_i is the substitution effect of marker i with distribution of $N\left(0, \sigma_{g_i}^2\right)$, N is the total number of markers, $\mathbf{e}$ is the vector of residual errors with distribution of $N\left(0, \mathbf{I}\sigma_e^2\right)$, $\mathbf{X}$ is the design matrix for $\mathbf{b}$, and $\mathbf{z}_i$ is a vector of indicators for genotypes of marker i with values equal to 0, 1, or −1 to indicate the marker genotypes 12, 22, and 11, respectively. The marker effect variance $\sigma_{g_i}^2$ was assumed a priori to be zero with a probability of π or to follow a scaled inverse chi-squared distribution, $\chi^{-2}(v, S^2)$, with a probability of $(1 - \pi)$ and parameter $v = 4.234$ and $S = 0.0429$ (Meuwissen et al., 2001). The prior distribution for the error variance, σ_e^2, was a scaled inversed chi-squared distribution with parameter $v = -2$ and $S = 0$. The exact ratio of number of simulated QTL to number of markers was taken as the value for $(1 - \pi)$. The Monte Carlo Markov Chain was run for 10,000 cycles with 100 cycles of Metropolis-Hastings sampling in each Gibbs sampling, and the first 2,000 cycles were discarded as burn-in. All the samples of marker effects from later cycles were averaged to obtain the estimates of marker effects.

Genomic Breeding Values Prediction

We compared 3 different approaches to predict GEBV using the LD marker panels BayesB, TABLUP, and GBLUP.

For TABLUP and GBLUP, the GEBV of all genotyped individuals were predicted by solving the mixed model equations based on the following model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \qquad [2]$$

where $\mathbf{u}$ is the vector of random polygenic effect of all individuals of both training and validation populations, which is the EBV in conventional BLUP and GEBV in TABLUP or GBLUP. For TABLUP, $\text{Var}(\mathbf{u}) = \mathbf{TA}\sigma_u^2$, where $\mathbf{TA}$ is a marker-based trait specific relationship matrix, which was constructed following the rule proposed by Zhang et al. (2010). In this study, the estimated marker effects by BayesB were used to weight markers in the $\mathbf{TA}$ matrix. For GBLUP, $\text{VAR}(\mathbf{u}) = \mathbf{G}\sigma_u^2$, where $\mathbf{G}$ is a marker-derived relationship matrix without weighting on any markers (VanRaden, 2008). The simulated variance components were provided to the mixed model equations.

For BayesB, the GEBV of a genotyped individual was calculated as the sum of all marker effects according to its marker genotypes (Meuwissen et al., 2001).

LD Panel Design

Two types of LD panels were designed for comparison: SLD and ELD panels. The markers in both types of LD panels were subsets of the HD panel with 10,000 simulated markers. To obtain an ELD panel, the first marker from each bin of markers in the HD panel was kept (e.g., the first of each 10 markers was kept to obtain an ELD panel with 1,000 markers). In this way, different ELD panels with 5,000, 2,000, 1,000, 500, or 200 markers were designed. For the SLD panels, the specific numbers of markers were selected based on the size of their estimated marker effects from BayesB in decreasing order. In this way, different SLD panels with 5,000, 2,000, 1,000, 500, 200, 100, or 50 markers were designed.

RESULTS

Accuracy of GEBV with HD Panels

The accuracies of GEBV in the validation population, expressed as correlation between GEBV and TBV, from BayesB, TABLUP, and GBLUP with HD panels under different heritabilities and effective population sizes are given in Tables 2 and 3, respectively. In all scenarios, TABLUP and BayesB performed very similarly and outperformed GBLUP. With the increase of N_{QTL} relative to M_e, the accuracies of BayesB and TABLUP decreased significantly, whereas the accuracies of GBLUP increased slightly, so that the advantage of BayesB and TABLUP over GBLUP decreased. In general, for all of the 3 methods, the accuracies of GEBV increased with the increase of the heritability and decreased with the increase of N_e.

Accuracy of GEBV with LD Panels in the Standard Scenario

Figure 2 shows the relative accuracies of GEBV, expressed as percentages of the accuracies of BayesB with an HD panel, of the 3 methods with SLD and ELD

Table 2. Accuracy of Bayesian variable selection method B (BayesB); BLUP with a trait-specific, marker-derived relationship matrix (TABLUP); and BLUP with a realized relationship matrix (GBLUP) with a high-density panel (10,000 markers) in the validation population under different heritabilities [effective population size (N_e) = 100]

		Heritability				
N_{QTL}[1]	Method	0.1	0.3	0.5	0.8	0.95
Low	BayesB	0.656 ± 0.045	0.848 ± 0.010	0.889 ± 0.011	0.936 ± 0.005	0.959 ± 0.003
	TABLUP	0.650 ± 0.042	0.842 ± 0.010	0.889 ± 0.009	0.938 ± 0.004	0.961 ± 0.003
	GBLUP	0.447 ± 0.016	0.598 ± 0.008	0.665 ± 0.007	0.756 ± 0.007	0.807 ± 0.007
Medium	BayesB	0.560 ± 0.028	0.737 ± 0.021	0.813 ± 0.013	0.892 ± 0.008	0.935 ± 0.005
	TABLUP	0.559 ± 0.029	0.736 ± 0.020	0.814 ± 0.013	0.894 ± 0.008	0.936 ± 0.005
	GBLUP	0.491 ± 0.021	0.616 ± 0.012	0.682 ± 0.009	0.768 ± 0.007	0.816 ± 0.005
High	BayesB	0.518 ± 0.022	0.694 ± 0.022	0.779 ± 0.015	0.874 ± 0.006	0.921 ± 0.004
	TABLUP	0.510 ± 0.018	0.691 ± 0.022	0.778 ± 0.014	0.872 ± 0.006	0.917 ± 0.004
	GBLUP	0.497 ± 0.018	0.625 ± 0.014	0.691 ± 0.013	0.775 ± 0.010	0.822 ± 0.008

[1] Number of simulated QTL; low: $N_{\text{QTL}} = 0.05M_e$, medium: $N_{\text{QTL}} = 0.3M_e$, and high: $N_{\text{QTL}} = 1M_e$ (where M_e = number of effective chromosome segments).

Table 3. Accuracy of Bayesian variable selection method B (BayesB); BLUP with a trait-specific, marker-derived relationship matrix (TABLUP); and BLUP with a realized relationship matrix (GBLUP) with a high-density panel (10,000 markers) in the validation population under different effective population sizes ($h^2 = 0.5$)

		Effective population size (N_e)				
N_{QTL}[1]	Method	50	100	200	500	1,000
Low	BayesB	0.928 ± 0.014	0.889 ± 0.011	0.865 ± 0.010	0.724 ± 0.020	0.611 ± 0.018
	TABLUP	0.925 ± 0.013	0.889 ± 0.009	0.865 ± 0.008	0.728 ± 0.017	0.621 ± 0.016
	GBLUP	0.682 ± 0.010	0.665 ± 0.007	0.633 ± 0.012	0.580 ± 0.015	0.551 ± 0.008
Medium	BayesB	0.868 ± 0.011	0.813 ± 0.013	0.754 ± 0.015	0.678 ± 0.011	0.571 ± 0.017
	TABLUP	0.866 ± 0.010	0.814 ± 0.013	0.750 ± 0.015	0.673 ± 0.011	0.560 ± 0.017
	GBLUP	0.703 ± 0.014	0.682 ± 0.013	0.636 ± 0.008	0.581 ± 0.009	0.545 ± 0.013
High	BayesB	0.833 ± 0.008	0.779 ± 0.015	0.702 ± 0.016	0.634 ± 0.014	0.577 ± 0.006
	TABLUP	0.832 ± 0.008	0.778 ± 0.014	0.699 ± 0.015	0.623 ± 0.016	0.565 ± 0.006
	GBLUP	0.723 ± 0.010	0.691 ± 0.013	0.625 ± 0.014	0.596 ± 0.011	0.558 ± 0.005

[1]Number of simulated QTL; low: $N_{\mathrm{QTL}} = 0.05M_e$, medium: $N_{\mathrm{QTL}} = 0.3M_e$, and high: $N_{\mathrm{QTL}} = 1M_e$ (where M_e = number of effective chromosome segments).

panels in the standard scenarios ($N_e = 100$, $h^2 = 0.5$). With an ELD panel, the relative accuracies decreased dramatically with the decrease of the number of markers. With an SLD panel, for BayesB and TABLUP, the relative accuracy decreased only slightly with the decrease of number of markers and more than 90% of the accuracy of BayesB with an HD panel could be retained, even when a very small proportion of total markers were used, no matter how many QTL controlled the trait of interest. On the contrary, for GBLUP, the relative accuracy increased with the decrease of number of markers, especially in the scenarios of low N_{QTL} relative to M_e. With both ELD and SLD panels, BayesB and TABLUP performed very similarly and outperformed GBLUP, and the advantage of BayesB and TABLUP over GBLUP decreased with the decrease of number of markers in the LD panels.

Effect of Heritability on the Accuracy of SLD Panel

The accuracies of TABLUP with an SLD panel relative to that of BayesB with an HD panel are plotted in Figure 3. The relative accuracies increased with the increase of heritability but decreased with the decrease of number of markers in the SLD panel. The lower the heritability, the faster the decrease of the accuracy along with the decrease of number of markers, especially in the case of high N_{QTL} relative to M_e. In the case of low N_{QTL} relative to M_e, the relative accuracies were more than 95% in all scenarios, whereas in the case of

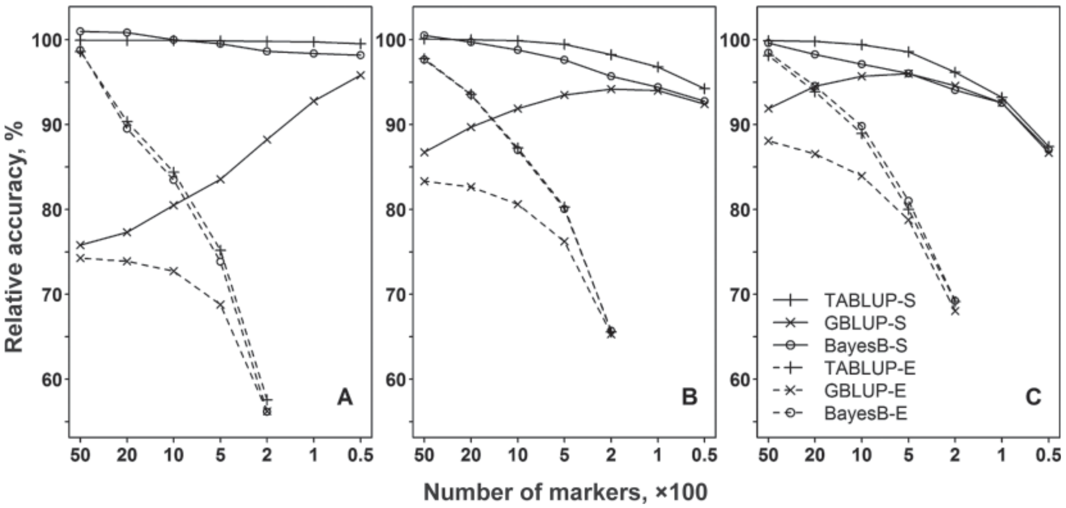

Figure 2. Relative accuracy of BLUP with a trait-specific, marker-derived relationship matrix (TABLUP); BLUP with a realized relationship matrix (GBLUP); and Bayesian variable selection method B (BayesB) with low-density (LD) panels in the standard scenario (effective population size = 100, heritability = 0.5). Accuracy is expressed as the percentage of the accuracy of BayesB with a high-density (HD) panel (10,000 markers). A: number of QTL (N_{QTL}) = $0.05M_e$, B: $N_{\mathrm{QTL}} = 0.3M_e$, C: $N_{\mathrm{QTL}} = 1M_e$ (where M_e = number of effective chromosome segments); S represents LD panel with selected markers (SLD), E represents LD panel with evenly spaced markers (ELD).

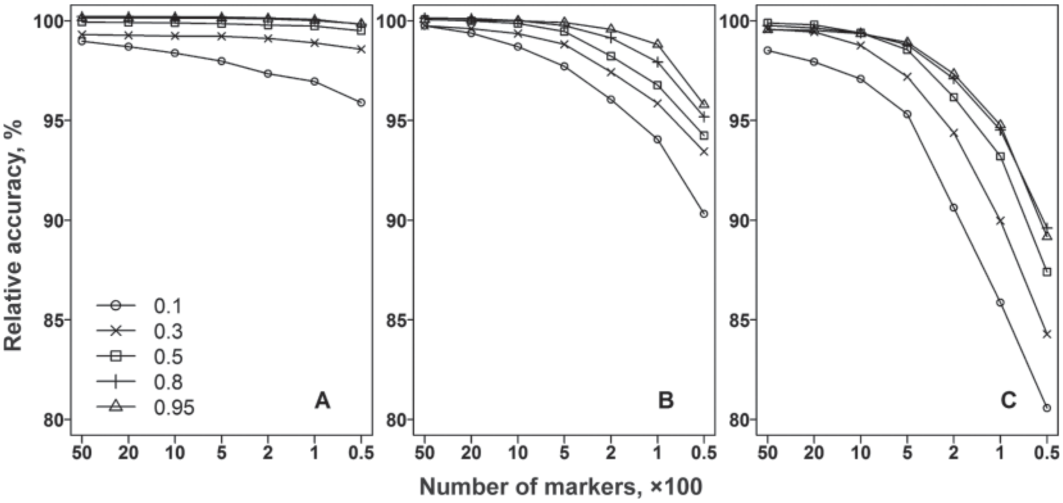

Figure 3. Effect of heritability on the relative accuracy of BLUP with a trait-specific, marker-derived relationship matrix (TABLUP) with selected low-density markers. Accuracy is expressed as percentage of the accuracy of Bayesian variable selection method B (BayesB) with a high-density (HD) panel (10,000 markers). Different lines represent different heritabilities. A: number of QTL (N_{QTL}) = $0.05M_e$, B: $N_{\mathrm{QLT}} = 0.3M_e$, C: $N_{\mathrm{QLT}} = 1M_e$ (where M_e = number of effective chromosome segments).

high N_{QTL} relative to M_e, the relative accuracies could drop down to below 90%. It should be noted that when the number of markers was over 200 or 500 (i.e., 2 or 5% of the total number of markers in the HD panel), over 90 or 95% of the accuracy of the HD panel could be retained by TABLUP with an SLD panel, no matter what the heritability and N_{QTL} relative to M_e.

Effect of Effective Population Size on the Accuracy of SLD Panels

The relative accuracies of TABLUP with an SLD panel under different N_e are shown in Figure 4. The relative accuracy decreased with the decrease of number of markers in the LD panels and with the increase of N_e. The larger the effective population size, the faster drop of the relative accuracy with the decrease of number of markers, especially in the case of high N_{QTL} relative to M_e. It should be noted that when the number of markers was over 500, over 90% of the accuracy of the HD panel could be retained by TABLUP with an SLD panel, no matter what the effective population size and N_{QTL} relative to M_e.

DISCUSSION

The main objective of this study was to investigate the accuracy of genomic selection with LD panels designed with different strategies under different numbers of QTL relative to number of effective chromosomal segments, heritabilities, and effective population size.

To implement genomic prediction with LD panels, breeders need to make the decision of using an ELD panel or an SLD panel in the selection population. The advantage of SLD panels over ELD panels was clear from the present study (Figure 2) and some other studies (Habier et al., 2009; Weigel et al., 2009; Cleveland et al., 2010). With the same number of markers, the loss of accuracy using an ELD panel will be larger than using an SLD panel. Moreover, the loss of accuracy with ELD panels becomes larger over generations (Habier et al., 2009). The markers in the SLD panel, however, are trait specific. If many target traits are selected simultaneously, the SLD panel needs to include the markers selected for all traits. Therefore, a trade-off exists between the number of traits included in a breeding program and the number of trait-specific markers for each trait included in the SLD panel. For instance, the results of this study showed that GS with the LD panel with 200 selected markers (2% of the total number of markers in the HD panel) would retain over 95% of the accuracy of GS with the HD panel in most scenarios. Suppose 5, statistically independent, target traits are to be selected in the breeding program and 200 markers are chosen for each trait, respectively. The SLD panel should contain 1,000 markers, which is equal to $1N_eL$ in this study, with $N_e = 100$ and the genome length $L = 10$ morgans. In dairy cattle, the actual length of the genome L is about 30 morgans, so an LD panel with 1,000 markers in this study is equivalent to an LD chip with 3,000 markers in dairy cattle (assuming $N_e = 100$). In comparison with an ELD panel with the same number of markers commonly used for the 5 traits, the accuracies of GS with the SLD panel are still higher than that with the ELD panel (Figure 2). If the traits are genetically correlated, some common markers

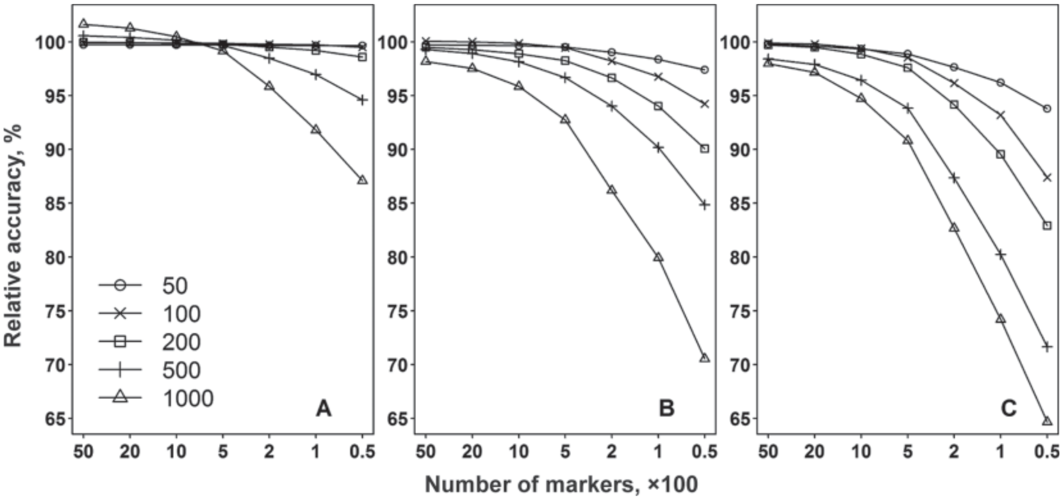

Figure 4. Effect of effective population size on the relative accuracy of BLUP with a trait-specific, marker-derived relationship matrix (TABLUP) with selected low-density markers. Accuracy is expressed as percentage of the accuracy obtained by Bayesian variable selection method B (BayesB) with a high-density (HD) panel (10,000 markers). Different lines represent different effective population sizes. A: number of QTL (N_{QTL}) = $0.05M_e$, B: $N_{\mathrm{QTL}} = 0.3M_e$, C: $N_{\mathrm{QTL}} = 1M_e$ (where M_e = number of effective chromosome segments).

with effects on multiple traits should exist, so the SLD can be designed to contain some common markers for multiple traits and some specific markers for each trait. This would further improve the performance of the SLD panels. If the number of target traits is very large (e.g., 20), the LD panel could be designed to be a mixture of evenly spaced markers with some trait-specific markers for those traits which deviate much from the infinitesimal model. Then, the overall predicting ability of the mixed ELD panel can be no worse than a pure ELD panel, whereas for those traits deviating much from the infinitesimal model, the predicting ability of the mixed ELD panel can be higher than a pure ELD panel. Therefore, the number of target traits in a breeding program and the genetic architecture of these traits decide the trade-off between SLD and ELD panels.

In this study, we did not use any linkage disequilibrium information of the HD panels in the training population when predicting GEBV in the validation population with LD panels. In other words, only the markers in the LD panels were used to predict GEBV. Actually, the genotypes of HD markers in the training population could be useful for the prediction of the GEBV of individuals in the selection population. Habier et al. (2009) and Cleveland et al. (2010) proposed a way to use the linkage disequilibrium information of the HD panel in the training population to infer the probabilities of chromosomal segments or missing genotypes in the LD panels in the selection population. Another way of using linkage disequilibrium information of an HD panel in the training population is to impute the missing genotypes in an LD panel. Weigel et al. (2010) demonstrated that the accuracy of imputation would be 0.70 to approximately 0.99, depending on how many markers are missing in the LD panel. The use of the linkage disequilibrium information of an HD panel could result in some differences in the gain of accuracy between the ELD and SLD panels. The SNP with the strongest effects will be directly genotyped in the SLD panels, whereas they may be imputed in the ELD panels.

According to Daetwyler et al. (2010), the accuracy of genomic prediction, r, is a function of the number of phenotypic records (N_p) in the training population, heritability (h^2) of the trait, N_{QTL}, and M_e, which can be estimated by $2N_eL/\ln(4N_eL)$ (Goddard, 2009), and can be predicted using the formula $r = \sqrt{N_ph^2 \big/ [N_ph^2 + min(N_{QTL}, M_e)]}$. This derivation is based on the assumption that the marker density is high enough to explain all genetic variance. Violation of this assumption is likely the main contributor to observed accuracies being lower than predicted. We compared the predicted accuracies using this formula and the observed ones in the simulation with HD panels in different scenarios (Table 4). The observed accuracies are generally lower than the predicted ones, particularly when N_{QTL} relative to M_e is low and h^2 is also low. Daetwyler et al. (2010) also obtained similar results and pointed out a need to improve the formula to predict the accuracy more accurately.

As shown in the results, the accuracy of GS decreased generally with the decrease of h^2, the increase of N_e, and the increase of N_{QTL} relative to M_e. One exception is that the accuracy of GBLUP was not affected by or

Table 4. Predicted accuracies and observed accuracies from Bayesian variable selection method B (BayesB) in different scenarios

		Predicted accuracy[2]			Observed accuracy		
		N_{QTL} relative to M_e			N_{QTL} relative to M_e		
N_e[1]	h^2	Low	Medium	High	Low	Medium	High
50	0.5	0.993	0.963	0.890	0.928	0.868	0.833
100	0.5	0.988	0.935	0.821	0.889	0.813	0.779
200	0.5	0.978	0.888	0.727	0.865	0.754	0.702
500	0.5	0.953	0.789	0.575	0.724	0.678	0.634
1,000	0.5	0.917	0.685	0.458	0.611	0.571	0.577
100	0.1	0.945	0.762	0.541	0.656	0.560	0.518
100	0.3	0.980	0.898	0.745	0.848	0.737	0.694
100	0.5	0.988	0.935	0.821	0.889	0.813	0.779
100	0.8	0.993	0.958	0.877	0.936	0.892	0.874
100	0.95	0.994	0.964	0.893	0.959	0.935	0.921

[1] Effective population size.

[2] Predicted accuracy was calculated from the formula $r = \sqrt{N_p h^2 \big/ [N_p h^2 + min(N_{QTL}, M_e)]}$, where r is the accuracy of genomic prediction, N_p is the number of phenotypic records in the training population, N_{QTL} is the number of simulated QTL, and M_e is the number of effective chromosome segments. Low: $N_{\mathrm{QLT}} = 0.05M_e$, medium: $N_{\mathrm{QLT}} = 0.3M_e$, high: $N_{\mathrm{QTL}} = 1M_e$.

even increased slightly with the increase of N_{QTL} relative to M_e (Tables 2 and 3; Figure 2). This is reasonable, as GBLUP is based on the infinitesimal model, which implies that the accuracy of GBLUP is not affected by the number of QTL. This phenomenon has been discussed in detail by Daetwyler et al. (2010). With LD panels, the relative accuracy of TABLUP decreased slightly with N_{QTL} relative to M_e in the standard scenario (Figure 2), but decreased quickly when h^2 was low (Figure 3) or N_e was high (Figure 4). This is because, in the case of low h^2 or high N_e, the accuracies of marker effect estimates will be low, unless more individuals and a denser HD panel are used in the training population.

Several studies have pointed out that the number of phenotypic records is an important factor affecting the performance of genomic prediction (Daetwyler et al., 2008; Goddard, 2009; Meuwissen, 2009; Daetwyler et al., 2010). In this study, the training population size N_p was fixed at 1,000 for all scenarios and the genome length L was fixed at 10 morgans. However, for N_e equal to 50, 100, 200, 500, and 1,000, this training population size was 2, 1, 0.5, 0.2, and 0.1 N_eL, respectively. Meuwissen (2009) showed that a high accuracy of prediction requires $2N_eL$ individuals in the training population. The training population size in this study represents a good to poor level of number of phenotypic records with respect to different N_e. For an N_e of 50, the training population size is $2N_eL$, and the results fit well with the expectation of Meuwissen (2009), not only for scenarios with HD panels (Table 3), but also for scenarios with SLD panels (Figure 4). On the other hand, for N_e of 1,000, the training population size is only $0.1N_eL$, leading to the bad accuracies of GEBV in scenarios with either HD or LD panels. An accurate estimate of genetic value with the LD panel needs a sufficient number of phenotypic records, but it does not appear to affect the ranking of different approaches in the present study.

Three methods, BayesB, GBLUP, and TABLUP, were compared for predicting GEBV with ELD and SLD panels in the scenario of $N_e = 100$ and $h^2 = 0.5$ (Figure 2). The BayesB and TABLUP methods performed almost identically and always outperformed GBLUP. Recently, Shepherd et al. (2010) presented an improved method to perform BayesB-type genomic selection estimates. Because many papers compare their method to the original BayesB (Meuwissen et al., 2001), we used that for comparison. However, we have tested the improved BayesB for some scenarios using the code provided by the authors. This showed that the performance for the scenarios that we studied ($h^2 = 0.5$, $N_e = 100$, No. markers = 10,000, size of reference population = 1,000), the improved BayesB performed slightly worse than the original BayesB. The advantage of BayesB and TABLUP over GBLUP increased with the increase in number of markers. This illustrated that BayesB benefits more from higher-density panels (Meuwissen, 2009). It should be noted that in the comparison, the exact ratio of number of simulated QTL to number of markers was applied to BayesB. This would result in an advantage of BayesB and TABLUP over GBLUP. However, the same relative performances of the 3 methods were also observed in other scenarios with different N_e and h^2 (data not shown). The advantage of TABLUP over GBLUP demonstrated that selecting a subset of markers and weighting them to construct a relationship matrix for the mixed model equations is a robust way to predict GEBV with LD panels, if good prior knowledge on marker effects for the trait of interest can be obtained.

CONCLUSIONS

To implement genomic selection with LD panels, a training population of sufficient size and genotyped with an HD panel is necessary. The trade-off between ELD and SLD must be considered, which depends on the number of target traits in a breeding program and the genetic architecture of these traits. For the ELD panel, the loss of accuracy is large and its persistency of predicting ability is potentially weak. For the SLD panel, the genetic architecture of trait of interest and the number of target traits in the breeding program should be taken into consideration. The TABLUP method was shown to perform well with an LD panel and should be explored further. Genomic selection with LD panels could be feasible and cost-effective, though a further detailed genetic and economic analysis is recommended before implementation.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 30800776, 30972092; Beijing), the State High-Tech Development Plan of China (Grant No. 2008AA101002), the National Key Basic Research Program of China (Grant No.2006CB102104), and the Innovation Fund for Graduate Student of China Agricultural University. D.-J. de Koning acknowledges support from the Biotechnology and Biological Sciences Research Council (BBSRC) through an Institute Strategic Program Grant to the Roslin Institute (Roslin, UK). We are grateful to the two anonymous reviewers for their insightful comments and constructive suggestions that greatly improved our manuscript.

REFERENCES

Cleveland, M. A., S. Forni, N. Deeb, and C. Maltecca. 2010. Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. BMC Proc. 4(Suppl 1):S6.

Coster, A., J. W. M. Bastiaansen, M. P. L. Calus, C. Maliepaard, and M. C. A. M. Bink. 2010. QTLMAS 2009: Simulated dataset. BMC Proc. 4(Suppl. 1):S3.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3:e3395.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longman Group Ltd., New York, NY.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257.

Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10:381–391.

González-Recio, O., D. Gianola, N. Long, K. A. Weigel, G. J. Rosa, and S. Avendaño. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. Genetics 178:2305–2313.

González-Recio, O., D. Gianola, G. J. Rosa, K. A. Weigel, and A. Kranis. 2009. Genome-assisted prediction of a quantitative trait measured in parents and progeny: Application to food conversion rate in chickens. Genet. Sel. Evol. 41:3.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182:343–353.

Harris, B. L., and W. A. Montgomerie. 2009. Current status of the use of genomic information in the national genetic evaluation in New Zealand. Pages 35–38 in Proc. of the Interbull international Workshop in Genomic information in genetic evaluations. Interbull, Uppsala, Sweden.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443.

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

Jacquard, A. 1974. The Genetic Structure of Populations. Springer-Verlag, New York, NY.

König, S., H. Simianer, and A. Willam. 2009. Economic evaluation of genomic breeding programs. J. Dairy Sci. 92:382–391.

Long, N., D. Gianola, G. J. Rosa, K. A. Weigel, and S. Avendano. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: Application to early mortality in broilers. J. Anim. Breed. Genet. 124:377–389.

Meuwissen, T. H. E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41:35.

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.

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

Shepherd, R. K., T. H. Meuwissen, and J. A. Woolliams. 2010. Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. BMC Bioinformatics 11:529.

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

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2010. Combining different marker densities in genomic evaluation. Proc. 2010 Interbull Meeting, Riga, Latvia. Interbull, Uppsala, Sweden.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, and D. B. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93:5942–5949.

Weigel, K. A., G. de los Campos, O. González-Recio, H. Naya, X. L. Wu, N. Long, G. J. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248–5257.

Weigel, K. A., C. P. Van Tassell, J. R. O'Connell, P. M. VanRaden, and G. R. Wiggans. 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93:2229–2238.

Zhang, Z., J. F. Liu, X. Ding, P. Bijma, D.-J. de Koning, and Q. Zhang. 2010. Best linear unbiased prediction of genomic breeding values using trait-specific marker-derived relationship matrix. PLoS ONE 5:e12648.