Asian-Aust. J. Anim. Sci.
Vol. 24, No. 2 : 162 - 172
February 2011

www.ajas.info

Predicting the Accuracy of Breeding Values Using High Density Genome Scans*

Deukhwan Lee and Daniel A. Vasco[1]**

Department of Animal Life and Environment Sciences, Hankyong National University Seokjeong-Dong 67, Ansung City, Kyeonggi-Do, 456-749, Korea

ABSTRACT : In this paper, simulation was used to determine accuracies of genomic breeding values for polygenic traits associated with many thousands of markers obtained from high density genome scans. The statistical approach was based upon stochastically simulating a pedigree with a specified base population and a specified set of population parameters including the effective and non-effective marker distances and generation time. For this population, marker and quantitative trait locus (QTL) genotypes were generated using either a single linkage group or multiple linkage group model. Single nucleotide polymorphism (SNP) was simulated for an entire bovine genome (except for the sex chromosome, n = 29) including linkage and recombination. Individuals drawn from the simulated population with specified marker and QTL genotypes were randomly mated to establish appropriate levels of linkage disequilibrium for ten generations. Phenotype and genomic SNP data sets were obtained from individuals starting after two generations. Genetic prediction was accomplished by statistically modeling the genomic relationship matrix and standard BLUP methods. The effect of the number of linkage groups was also investigated to determine its influence on the accuracy of breeding values for genomic selection. When using high density scan data (0.08 cM marker distance), accuracies of breeding values on juveniles were obtained of 0.60 and 0.82, for a low heritable trait (0.10) and high heritable trait (0.50), respectively, in the single linkage group model. Estimates of 0.38 and 0.60 were obtained for the same cases in the multiple linkage group models. Unexpectedly, use of BLUP regression methods across many chromosomes was found to give rise to reduced accuracy in breeding value determination. The reasons for this remain a target for further research, but the role of Mendelian sampling may play a fundamental role in producing this effect. (**Key Words :** Simulation-based Inference, Prediction Accuracy, Breeding Value, High Density Genome Scan, Single Nucleotide Polymorphism)

INTRODUCTION

A major goal of quantitative genetics involves the estimation of breeding values using high density molecular data. Using the predictive power of quantitative genetic theory, in principle it is now possible, using novel and efficient computational approaches (VanRaden, 2008), to predict breeding value accuracy for a high density genome scan. For example, whole genomic selection for animal breeding is a rapidly advancing technology for guiding selection of superior breeding stock in the livestock industries. Using high density whole genome scanning technologies, it is now possible to assay up to 50K single nucleotide polymorphism (SNP) loci per animal in thousands of animals (Van Tassell et al., 2008; Matukumalli et al., 2009). Using these assays to develop genetic prediction models, it may be possible to improve the rate of genetic gain by reducing the generation interval without any loss in accuracy using a technology which has become known as genomic selection (Meuwissen et al., 2001). The technology has led to considerable excitement for applications in commercial and industrial settings (Schaeffer, 2006).

Recent studies have demonstrated that genomic selection was better than selection based upon best linear unbiased prediction (BLUP) using pedigree information in terms of the accuracy of predicted breeding values (Villanueva et al., 2005; VanRaden, 2008; Calus et al., 2008; Solberg et al., 2008; VanRaden et al., 2009) in simulated and actual data. VanRaden (2008) developed

* This Study was supported by Technology Development Program for Agriculture and Forestry, Ministry for Agriculture, Forestry and Fisheries, Republic of Korea.

** Corresponding Author : Deukhwan Lee. Tel: +82-31-670-5091, Fax: +82-31-676-5091, E-mail: dhlee@hknu.ac.kr

[1] Animal Genomics, Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA.

Received April 26, 2010; Accepted September 19, 2010

computing methods for the estimation of genomic predictions (linear or nonlinear prediction) for dairy cattle using whole genome high density mapping data produced with a 50K SNP chip and assuming chromosomes of equal length. Calus et al. (2008) also investigated the accuracy of breeding values produced by genomic selection in a simulation study using different map densities and haplotype structures in a 3 Morgan genome and within an outbred and unselected population. They found that the greatest benefit of genomic selection with high density genomic information was for low heritability traits. Solberg et al. (2008) investigated the relationship between marker density and effective population size on the accuracy of breeding values predicted using genomic selection.

The objective of this study was to investigate the effect of the number of linkage groups on accuracy of breeding values using genomic selection. Our approach may be used to predict the accuracy of molecular breeding values by calibrating the SNP density used in the simulation with respect to the extent of linkage disequilibrium found in cattle.

MATERIALS AND METHODS

Simulation method

It was assumed that SNP marker loci were evenly distributed and QTL loci were randomly scatter distributed. Two scenarios were used for the statistical modeling of the distribution of SNP and QTL genotypes. The first scenario assumed a single linkage group model (SLM) in which the organism's genome was comprised of a single chromosome of 1 M length. In the second scenario, a multiple linkage group model (MLM) was considered, consisting of 29 autosomes, to infer the relationship between independent linkage groups and the amount of influence of marker information on the estimation of genomic breeding values. Lengths of each chromosome are shown in Table 1. Both scenarios assumed that there was no mutation. Genotypes in the founder (base) population were in Hardy-Weinberg equilibrium and at gametogenesis in each subsequent generation such that maternal and paternal chromosome segments were inherited assuming no interference in recombination and utilizing Haldane's map distances (Lynch and Walsh, 1998). Recombination events were simulated under a Poisson distribution and were positioned at random on each chromosome. Linkage disequilibrium between markers was estimated by computing r^2 values (Liu, 1998).

Pedigree

A founder population was created consisting of 1,000 animals with even sex ratio. From this chromosomal inheritance was simulated beginning with the random sampling of a founder population of 50 sires. Each sire was mated at random to 10 dams. Each dam randomly produced 1-3 progeny, each having an equal chance of being male or female. Using this method, 1,000 individuals were propagated each generation. Starting from the founder population, 10 generations were created by this method to produce approximately 11,000 individuals. Genealogical information was tracked for all individuals in the population allowing the transmission of gametes and recombination events to be recorded at each generation. In this respect, our forward stochastic simulation method is similar to that of Libiger and Schork (2007) who utilized the approach to study chromosome segment sharing among a group of arbitrarily related individuals. In this paper, a similar method was used to study the effects of chromosomal segments which were identical by descent (IBD) on the accuracy of genomic prediction. Similar stochastic modeling approaches have recently been used to study complex diseases (Peng and Kimmel, 2005; Peng et al., 2007) and the backward simulation of ancestors of sampled individuals (Gasbarra et al., 2005). The underlying distributions for simulating genealogies of sampled populations are complex, therefore simulation-based methods are a powerful tool for statistically modeling this complexity (Guttorp, 1995).

Single linkage group model

This model, which simulated a 1 M (Morgan) chromosome having 33 biallelic QTLs, was genotyped with sufficient SNP to produce SNP spacing intervals of 0.08 cM, 0.20 cM, 0.30 cM and 1.00 cM. QTLs were positioned at random along the chromosome; marker loci were also randomly distributed and none of the markers directly affected any trait. A total of 1,250, 500, 333 and 100 biallelic marker loci were assigned to the chromosome for all founder individuals.

In the founder population, the allele frequency of each marker and QTL was assigned from uniform distribution within a range of 0.02 to 0.98. The allele substitution effect x at each QTL locus was generated from a Gamma distribution. This assumed that the additive effect at each QTL locus i was determined by the density function (see example Wu et al., 2007, p. 19):

$$f(x) = \frac{\alpha^{\beta} e^{-\alpha x} a^{\beta-1}}{\Gamma(\beta)}$$

where α is the scale parameter and β is the corresponding shape parameter which were defined as α = 1.66 and β = 0.40. This density allowed the modeling of a highly skewed distribution of QTL effects with equal probability of positive or negative effects (Hayes and

Table 1. Length of linkage groups and the number of marker and QTL loci for the SLM and MLM scenarios (unit = centi-morgans)

	Length	Number of markers				Number of QTLs
		0.08 cM	0.20 cM	0.30 cM	1.00 cM	3.50 cM[1]
Single linkage group						
	100.0	1,250	500	333	100	33
Multiple linkage group						
BTA1	173.9	2,173	869	579	173	49
BTA2	141.5	1,768	707	471	141	40
BTA3	183.2	2,290	916	610	183	52
BTA4	127.8	1,597	639	425	127	36
BTA5	137.6	1,720	688	458	137	39
BTA6	157.7	1,971	788	525	157	45
BTA7	147.3	1,841	736	490	147	42
BTA8	183.6	2,295	918	611	183	52
BTA9	122.6	1,532	613	408	122	35
BTA10	157.2	1,965	786	523	157	44
BTA11	150.9	1,886	754	502	150	43
BTA12	125.4	1,567	626	417	125	35
BTA13	146.0	1,825	730	486	146	41
BTA14	100.8	1,260	504	335	100	28
BTA15	85.8	1,072	429	285	85	24
BTA16	114.9	1,436	574	382	114	32
BTA17	138.0	1,725	690	459	138	39
BTA18	146.1	1,826	730	486	146	41
BTA19	98.8	1,235	494	329	98	28
BTA20	98.2	1,227	491	327	98	28
BTA21	105.6	1,320	528	351	105	30
BTA22	104.2	1,302	521	347	104	29
BTA23	75.9	948	379	252	75	21
BTA24	67.7	846	338	225	67	19
BTA25	99.3	1,241	496	330	99	28
BTA26	90.4	1,130	452	301	90	25
BTA27	87.1	1,088	435	290	87	24
BTA28	92.2	1,152	461	307	92	26
BTA29	71.0	887	355	236	71	20
Total	3,530.7	44,125	17,647	11,747	3,517	995

[1] Average map distance between flanking QTL loci was 3.0 cM for the single linkage group model.

Goddard, 2001). From this distribution, the largest QTL effect explained approximately 10% of the additive genetic variation in the trait. The simulated additive genetic variance at each locus was computed using the formula $\sigma^2_{g_i} = 2p(1-p)x^2$, where p is the allele frequency and x is the allele substitution effect (Falconer and Mackay, 1996). The total additive genetic variance was obtained by summing the variance components across all QTL. We also computed the breeding value of each animal as the summation of breeding values for each QTL without any dominance and epistatic effects. Environmental effects were modeled for individuals in the 8th and 9th generations assuming a normal distribution and three different heritabilities of 0.10, 0.30 and 0.50. Phenotype data were assigned only to these individuals (phenotyped individuals). Individuals on the 10th generation (juveniles) were assigned only genotypes but not phenotypes.

Multiple linkage group model

For the second scenario, the pedigree for the population was tracked as described above with genotype on every individual, but assumed 29 autosomes of estimated length by Barendse et al. (1997) and ignored sex chromosomes. Cattle were chosen to benchmark the method with distributions of markers on an observed genome. Bovine genomes approach sizes of 35.307 Morgans (Barendse et al., 1997) which was chosen as a rough benchmark for this

model. Markers were distributed approximately uniformly along chromosomes but at different densities (Table 1). The evenly spaced distances between flanking markers were assumed to be 0.08, 0.20, 0.30, and 1.00 cM for the multiple and single linkage group models. The average distance between QTLs was assumed to be 3.5 cM. The total number of genotyped SNP for the multiple linkage group case was 44,125, 17,647, 11,747 and 3,517 giving average intermarker spacings of 0.08, 0.20, 0.30, and 1.00 cM, respectively. The total number of QTLs was 995 for the MLM simulations which was considered a realistic representation. QTL effects were assigned as before in the case of the SLM and the variances of the allele substitution effects on each QTL locus and total genetic variance were calculated.

Genetic prediction

Two relationship matrices were used for genetic prediction: the genomic relationship matrix (GRM) and the pedigree-based relationship matrix (PRM). A GRM was computed using SNP data for individuals from generations 8 to 10. Allele frequencies were estimated for individuals from generations 8 to 10, but it was assumed that allele frequencies were known for all other generations. The genomic relationship matrix, *G*, calculated as in VanRaden (2008) was:

$$G = \frac{WW'}{2\sum_j p_j(1-p_j)}$$

where $W = M-P$, M is n (number of animals) x m (number of SNP) contains elements -1, 0, and 1 respectively, for homozygote, heterozygote and alternate homozygote classes, P is a matrix comprising columns of allele frequencies scaled from -1 to 1, such that column j of P contains the element $2(p_j - 0.5)$, and p_j is the frequency of the second allele at locus j.

For PRM, the numerator relationship matrix, *A*, was computed using the recursive algorithm proposed by Aguilar and Misztal (2008) (which was based on an earlier algorithm by Quaas (1976)). This was done for the whole population starting from the founder population to the final generation.

Since we assumed that individuals from generations 8 to 9 were measured for both genomic information and phenotypes, genomic breeding values (GBV) of animals from generation 10 could be estimated using phenotype and genotype information from their ancestors. GBV's were estimated using the prediction equations presented above. Likewise, breeding values using pedigree information (PBV) were estimated for all individuals in generation 8 and 9. A linear mixed model for each animal corresponding to each set of observations was created using $y_i = a_i + e_i$ where $a \sim N(0, G\sigma_a^2)$ and $e \sim N(0, I\sigma_e^2)$. For breeding value estimation utilizing genomic information, this model may be modified as $y_i = \sum_{j=1}^{m} W_{ij}u_j + e_i$, where u_{ij} is the j[th] marker effect for the i[th] animal such that breeding values may be obtained by summing across all marker loci $(\sum_{j=1}^{m} W_{ij}u_j)$ (see also VanRaden (2008)). With the assumption of normality for random effects for breeding values and residuals, the estimated breeding values can be obtained using selection index equations such as $\hat{a} = \frac{\text{cov}(a,y)}{\text{var}(y)}y$. This yields:

$$\hat{a} = G(G + I\lambda)^{-1}y \quad \text{where} \quad \lambda = \frac{\sigma_e^2}{\sigma_a^2}$$

For predicting breeding values for juveniles without observation, this equation may be modified as: $\hat{a} = C(G + I\lambda)^{-1}y$, where $C = \frac{W_2W'}{2\sum_j p_j(1-p_j)}$ and W_2 is the genotypic information matrix of juvenile population.

The breeding value estimates produced using pedigree information (PBV) were obtained using the standard mixed model $y_i = a_i + e_i$ assuming $a \sim N(0, A\sigma_a^2)$ and $e \sim N(0, I\sigma_e^2)$, where A is the numerator relationship matrix (defined above).

Accuracy of breeding value estimates are represented as the correlation between true breeding values and estimated breeding values and 1,000 replicates were processed for each case to estimate the accuracy.

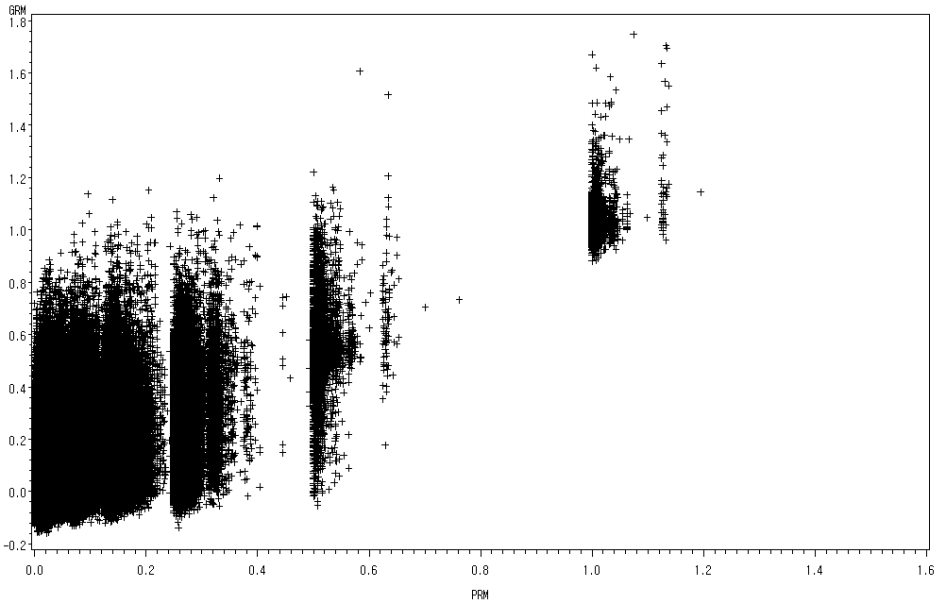
Validation of linkage disequilibrium

For validating the simulation in terms of genealogical processes exhibiting recombination at the gematogenesis step in the population, the pedigree generation was extended up to 1,000. The same assumption was made as used in the genomic prediction step except a point mutation rate of 2.5×10^{-3} was assumed until generation 800. Linkage disequilibrium between flanking markers was computed in terms of standard r^2 values (Liu, 1998).

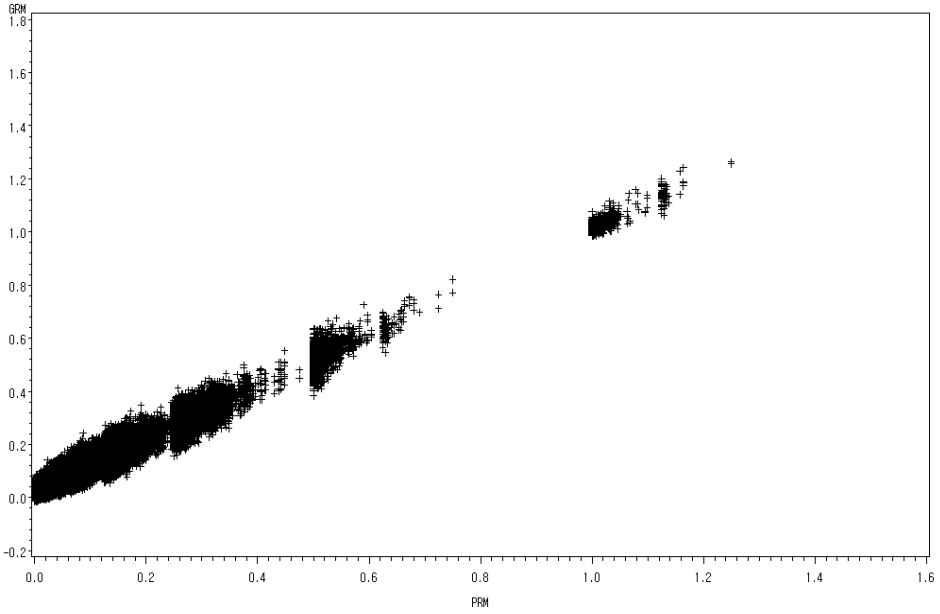
RESULTS

Efficiency of inbreeding coefficients

Due to the absence of selection and negligible effects of



(A)



(B)

Figure 1. Plot of the genomic relationship matrix versus the pedigree relationship matrix with 0.08 cM condition using a single linkage group model (A) and a multiple linkage group model (B).

drift, allele frequency estimates obtained in generation 8 to 10 were highly correlated with the true allele frequencies from the founder population (r = 0.97) for all models (SLM or MLM) regardless of marker density.

The genomic relationship coefficients between relatives and inbreeding coefficients of each animal were more variable than their expectations from pedigree information (PRM) for the SLM. However, for the MLM the difference between GRM and PRM was much reduced and the two were highly correlated (Figure 1A and 1B, Table 2). Average inbreeding coefficients increased by the same amounts when computed using the GRM or PRM as the number of generations increased in the simulation (see Table 3). As more genotype information is included in the sample, the average Mendelian sampling contribution effects are expected to become more centralized. Hence, with greater information (more SNP) in the sample, it is predicted that the average Mendelian sampling effects will produce a tightly centered distribution around the expected relationship between individuals in a population.

Table 2. Correlation and regression coefficients for genomic and pedigree relationship matrices as a function of map distances and linkage models for individuals from generation 8-10

	SLM		MLM	
	$\hat{r}$	$\hat{b}$	$\hat{r}$	$\hat{b}$
1.00 cM	0.34	0.97	0.90	0.99
0.30 cM	0.45	1.00	0.95	0.99
0.20 cM	0.49	0.99	0.95	0.99
0.08 cM	0.51	1.01	0.96	0.99

SLM = Single linkage group model, MLM = Multiple linkage group model, $\hat{r}$ = Estimates of correlation coefficients, $\hat{b}$ = Estimates of regression coefficients.

Table 3. Means and standard deviations of inbreeding coefficients using genomic information with different marker densities and pedigree for individuals from generation 8-10

Generation	Genomic inbreeding coefficient				Pedigree inbreeding coefficient
	0.08 cM	0.20 cM	0.30 cM	1.00 cM	
8	1.76±0.52	1.95±0.65	2.01±1.11	2.74±2.51	1.93±0.10
9	2.20±0.54	2.22±0.77	2.29±1.13	3.07±2.63	2.20±0.12
10	2.36±0.74	2.57±0.99	2.62±1.16	3.17±2.80	2.48±0.17

Genetic prediction accuracy for the SLM

For the low heritability model ($h^2 = 0.10$), the breeding values estimated using genomic information (GEBV) were generally more highly correlated to the true breeding values than those estimated using pedigree information (PEBV). The exception was for low marker densities (>0.30 cM) for both older individuals with phenotypes and juveniles without phenotypes (Figure 2A-1). The accuracies of GEBV using the estimated allele frequencies at generation 8 to 10 were almost the same as those using the true allele frequencies from the founder population regardless of map distances. For the 0.08 cM density model, accuracies of breeding values were 0.66 and 0.35 using genomic information and pedigree information, respectively, on individuals from generation 8 and 9. The estimates on juveniles were 0.58 and 0.35, respectively (Figure 2A-1). The correlation between GEBV and PEBV were estimated at 0.69 and 0.51 (Figure 3A).

In the case of the medium heritability model ($h^2 = 0.30$), prediction accuracies using both genomic information and pedigree information were higher than those obtained for the low heritability case (Figure 2B-1). Furthermore, genomic breeding values were estimated with higher accuracy than pedigree breeding values. For example, prediction accuracies with the 0.08 cM map density were 0.82 and 0.68 for GEBV and PEBV on phenotyped animals and 0.75 and 0.48 on juvenile individuals, respectively (Figure 2B-1). Correlation between GEBV and PEBV was 0.78 and 0.58 on phenotyped and juvenile individuals, respectively (Figure 3B). In comparing the GEBV and PEBV prediction accuracies for the 0.20 cM marker density, we found accuracies of 7% and 17% for GEBV which were significantly higher than those obtained with PEBV. Further, the deviations of accuracies (%) for GEBV from PEBV were always higher in juveniles than in phenotyped individuals.

As with previous results presented, in the high heritability model ($h^2 = 0.50$) with the 0.08 cM map density, GEBV accuracies were 0.88 and 0.82 for phenotyped and juvenile individuals, respectively (Figure 2C-1). These estimates were considerably greater than the accuracies obtained for PEBV of 0.78 and 0.55.

Genetic prediction accuracy for the MLM

When compared with the accuracies on the single linkage group model, the accuracies of GEBV for the MLM were lower than those in the SLM. The accuracies of GEBV were investigated by comparing GEBV with PEBV estimates for the MLM. Again, the estimated allele frequencies from generations 8-10 were almost the same regardless of map density and therefore the GEBV computed using estimated allele frequencies were very similar to GEBV estimated using true allele frequencies from the founder population. We investigated the accuracies of the GEBV using the estimated allele frequencies by comparing PEBV estimates for different map densities.

In the Low heritability model ($h^2 = 0.10$), breeding value estimates had prediction accuracies of 0.51 and 0.38 for phenotyped and juvenile individuals, respectively (assuming a high density marker map-0.08 cM)-see panel Figure 2A-2. These accuracies were higher than when pedigree information was used (0.48 and 0.35). Otherwise, lower accuracies were observed when compared to the SLM using the corresponding heritability and marker densities-

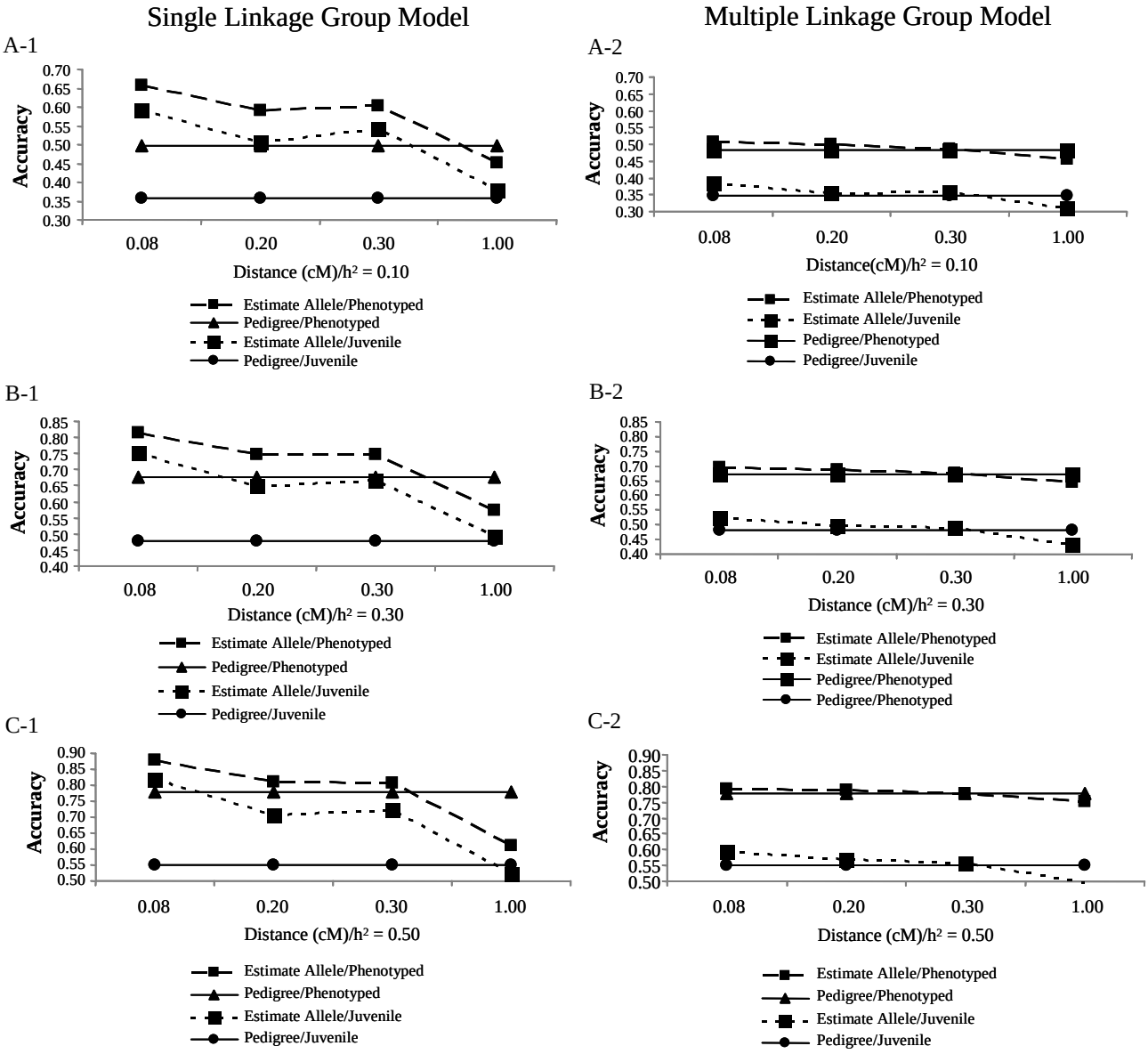


Figure 2. Plots for correlations between true breeding values and estimated breeding values under different map distances (cM) and heritabilities on the single linkage group and the multiple linkage group models.

See Figure 2. The accuracies using the 0.30 cM map density were similar to PEBV accuracies. Lower density maps are expected to give lower accuracies than those computed using PEBV. Such results are expected to be similar to those obtained for the single linkage group model.

In the model of Medium heritability (h^2 = 0.30), GEBV accuracies for a map density of 0.08 cM were slightly higher than the PEBV accuracies for phenotyped indviduals (0.69 vs. 0.67) and for juveniles (0.52 vs. 0.48). At a map density of 1.00 cM, accuracies of GEBV were lower than PEBV for phenotyped individuals (0.65 vs. 0.69) and juveniles (0.43 vs. 0.48) (Figure 2B-2).

Furthermore, in the High heritability case (h^2 = 0.50), trends were similar to those for the medium heritability case. GEBV accuracy for a map density of 0.08 cM density was 0.79 for phenotyped individuals and 0.60 for juveniles with comparable PEBV accuracy (0.78 and 0.55) (Figure 2C-2).

Influence of several linkage groups on accuracy of genomic breeding values

Table 4 shows the increase in accuracies of GEBV over PEBV for phenotyped and juvenile populations. For the SLM, the accuracy of breeding values using genomic information was greater than when using pedigree information. When we assumed 44K SNPs over 29 autosomes with 0.08 cM average density, the accuracy of breeding for juvenile was 3-5% higher than those using pedigree information. These results demonstrate that the

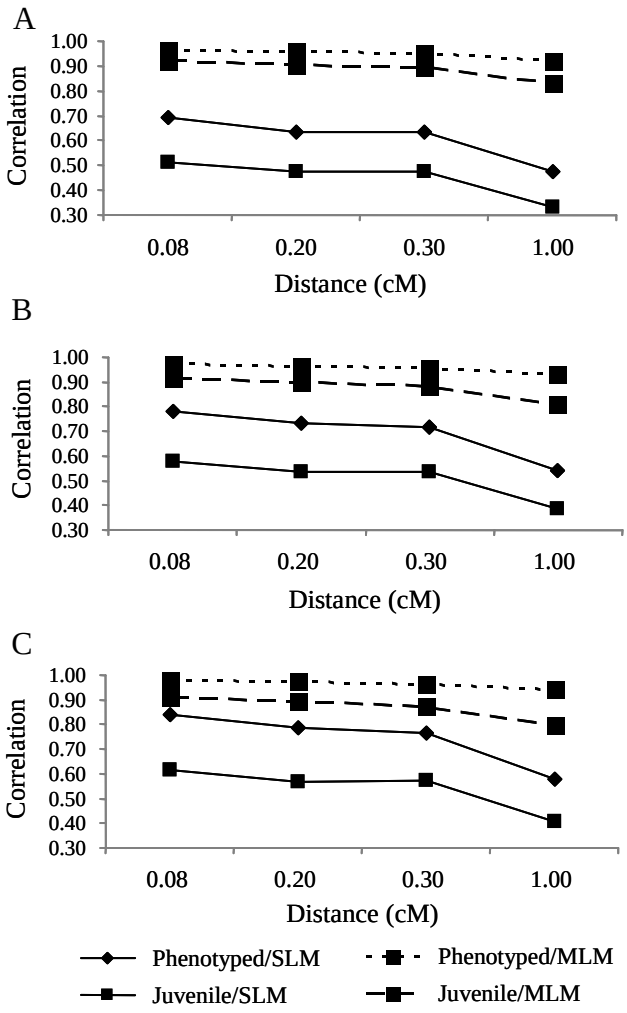


Figure 3. Plots for correlations between estimated breeding values using genomic information and estimated breeding values using pedigree information under several map distances (cM) and heritability of 0.10(A), 0.30(B), and 0.50(C) in the different linkage groups models (SLM vs. MLM).

accuracy for breeding value estimation might be influenced by the effects of Mendelian sampling over the QTL effects across chromosomes.

DISCUSSION

Accuracy of genetic prediction

For single linkage group models, the accuracy of genetic prediction was mainly influenced by heritability of the trait. Our results for the accuracies in the low heritability model were lower than equivalent results obtained by Solberg et al. (2008). We believe that this is because of the lower LD which may be expected to occur for expanding populations. Furthermore, deviations of accuracies were generally large for low density scans. This occurs due to weak linkage disequilibrium between marker loci. This result is in agreement with results of Calus et al. (2008). Furthermore, in the medium heritability model, the accuracies of GEBV in the 1.00 cM density model were similar or lower than those for PEBV. These results may be caused by high LD, as Calus et al. (2008) have recently argued. In our study, the linkage phase was assumed to be unknown, and thus accuracy of breeding values were generally lower than those reported by Calus et al. (2008) due to lower LD. In the high heritability model, the accuracy of GEBV for juvenile individuals (0.82) was similar to the estimate (0.85) obtained by Meuwissen et al. (2001) for a heritability of 0.50. The 0.08 cM map density in our simulation model corresponds to the 6.8/morgan model used by Calus et al. (2008) in their model. This is because the effective population size (*Ne*) in our simulation was about 182. Calus et al. (2008) obtained accuracies for 4/morgan and 8/morgan of 0.84 and 0.86, respectively, for juveniles. This is slightly higher than our result (0.82).

Comparing results on the single linkage group model, in the case of low density mapping these accuracies were more variable (possibly reflecting higher stochastic variation in Mendelian sampling present in simulation). This is likely to be related to the magnitude of LD that was present at the time the breeding value was estimated. Solberg et al. (2008) found that if the effective population size was doubled from 100, then twice the density of mapping was needed to obtain the same accuracy of breeding values. They found that the accuracy of breeding values for juveniles was about

Table 4. Deviations of accuracies (%) of the estimated genomic breeding values from the accuracies (%) of the estimated breeding values by pedigree information for phenotyped and juvenile individuals with different heritabilities and linkage groups

Individual	Distance	SLM[1]			MLM[2]		
		h^2 = 0.10	h^2 = 0.30	h^2 = 0.50	h^2 = 0.10	h^2 = 0.30	h^2 = 0.50
Phenotyped	0.08	17	14	10	2	2	2
	0.20	10	7	3	1	1	1
	0.30	12	7	3	0	0	0
	1.00	-4	-10	-17	-3	-3	-2
Juvenile	0.08	24	27	27	3	4	5
	0.20	15	17	16	0	2	2
	0.30	19	19	17	1	1	1
	1.00	3	1	-3	-4	-5	-5

[1] SLM = Single linkage group model. [2] MLM = Multiple linkage group model.

0.68 with a heritability of 0.50. We found that the accuracy of estimated breeding values for juveniles was 0.60. Solberg et al. (2008) estimated that about ~24,000 SNPs were needed for a 30 Morgan genome and effective population size (*Ne*) of 100. Our results showed that about 50k SNPs are needed to for a *Ne* of 182.

The degree of linkage disequilibrium (LD)

Macleod et al. (2006) argued that the power to detect a QTL explaining 5% of phenotypic variance with a marker having r^2 of 0.10 was about 0.80 on a population size of 2000. However, it is likely that there exist many QTLs having small effects on a given genome scan. Meuwissen et al. (2001) showed that the level of LD between adjacent markers should be $r^2 \geq 0.20$ to obtain reliable breeding values. Solberg et al. (2008) also found a drop in accuracy of 20% as marker spacing was increased from one marker every 0.5 cM to one marker every 4 cM on a population with Ne = 100. In realistic situations, LD will depend on mutation and random drift in the growing phase of the population. Our initial computational results did not consider mutation because of simplicity and because mutation is less important over the short time scales over which many animals are sampled in livestock studies. Also, we wanted to investigate breeding value accuracies using a scenario as close as possible to VanRaden (2008). However, we performed a preliminary study on the role of LD in a growing population in mutation-drift equilibrium. Figure 4 shows LD in this population at generation 1004. During the growing phase, mutations accumulated until generation 800. These occurred with a rate of 2.5×10^{-3} on each marker. The average r^2 for the population at generation 1004 was slightly over 0.2, when mutations were included, until generation 800. This estimate was similar to the report of DeRoss et al. (2008). They found that the average r^2 was greater than 0.20 between adjacent markers with a distance of 100 kb when using an Australian data set that was comprised of 383 Hostein-Friesian and 379 Angus animals. Thus, our results on breeding value accuracies are similar to those of other studies explicitly incorporating the effect of LD (for example, Calus et al., 2008).

Role of the distribution of QTL effects and the number of QTLs

Meuwissen et al. (2001) assumed that the distribution of QTL effects may be characterized as a Gamma distribution. Assuming the Gamma distribution of QTL requires a minimum of 100 or more QTL throughout the genome affecting a particular quantitative trait locus (Hayes et al., 2006). Even though VanRaden et al. (2008) simulated QTL effects from a weighted normal distribution, his estimates for breeding values were more highly correlated than the true breeding values in a nonlinear than a linear regression model. Because of the putative Gamma distribution

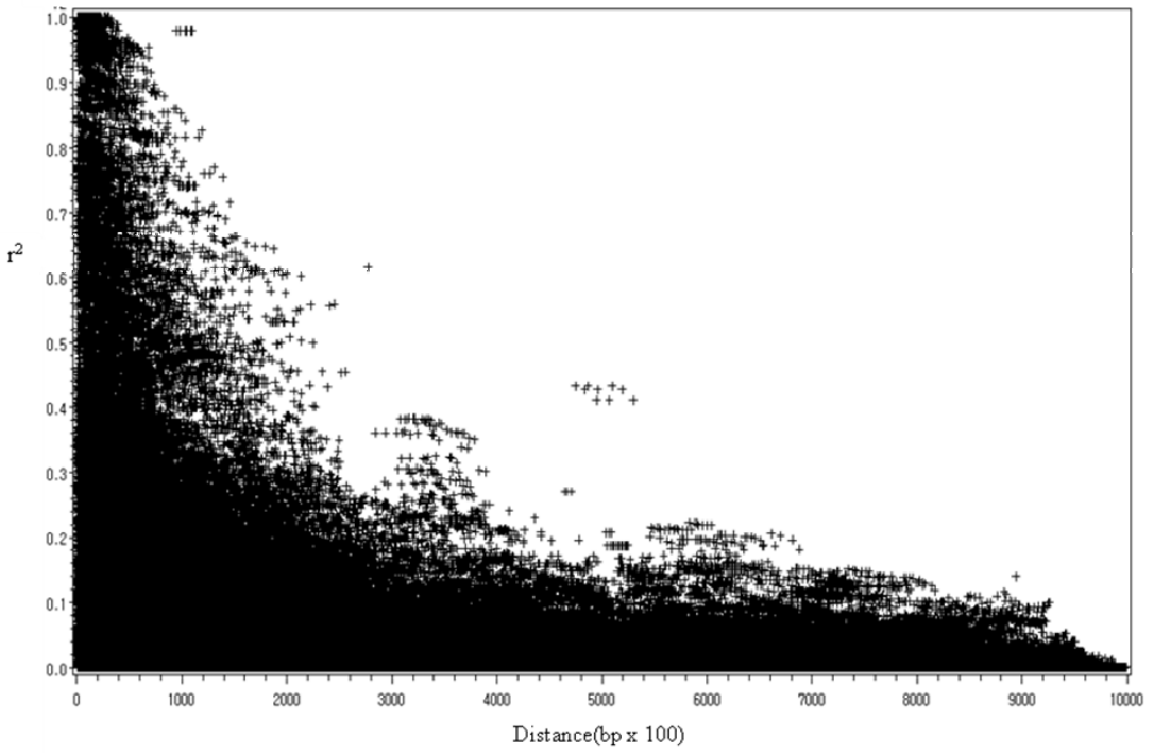


Figure 4. Plot of linkage disequilibrium between flanking markers in an expanding population sampled at generation 1004, with a mutation rate of 2.5×10^{-3} occurring until generation 800.

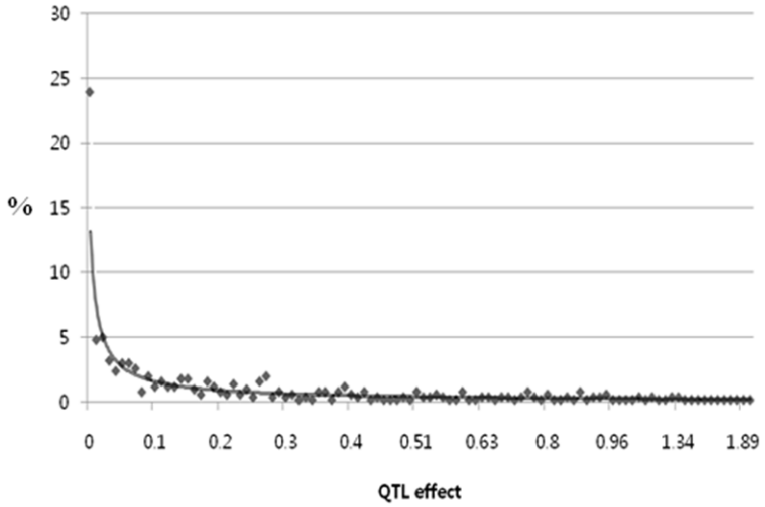


Figure 5. Proportion (%) with respect to the magnitude of QTL effect using the multiple linkage group model.

property of QTL effects, Meuwissen et al. (2001) suggested a Metropolis hashing algorithm for genomic selection. The QTL effects in our simulation data in the case of multiple linkage group models were distributed as shown in Figure 5. The total number of QTLs as well as each QTL effect will influence the accuracy of breeding values. However, we have not yet investigated the number of QTL effects on determining accuracy of breeding values.

Comparison between SLM and MLM

Figure 1 and Table 2 compare the effect of individual relatedness in determining breeding value accuracy when using genomic information versus those methods which utilize pedigree information. The deviation of genomic relatedness from half-sib or ful-sib relationships is dependent upon the number of markers used (VanRaden, 2007). We found that relatedness measures utilizing genomic information on our SLM were similar to those found by VanRaden (2007). However, in the case of our MLM in a computational methodology, those differences in attaining breeding value accuracy using the different measures are lost and the correlation between GRM and PRM approaches unity. The gist of this is that the overall accuracy of breeding values on MLM is reduced. The apparent reason for this loss of accuracy is that genomic information is being utilized inefficiently due to the effect of Mendelian inheritance across several chromosomes.

ACKNOWLEDGMENT

This research was supported by Technology Development Program for Agriculture and Forestry, Ministry for Food, Agriculture, Forestry and Fisheries, Republic of Korea.

REFERENCES

Aguitar, I. and I. Misztal. 2008. Technical note: Recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. J. Dairy Sci. 91:1669-1672.

Bailey, N. T. J. 1961. Introduction to the Mathematical Theory of Genetic Linkage. Oxford Universisty press.

Barendse, W., D. Vaiman, S. J. Kemp, Y. Sugimoto, S. M. Armitage, J. L. Williams, H. S. Sun, A. Eggen, M. Agaba, S. A. Aleyasin, M. Band, M. D. Bishop, J. Buitkamp, K. Byrne, F. Collins, L. Cooper, W. Coppettiers, B. Denys, R. D. Drinkwater, K. Easterday, C. Elduque, S. Ennis, G. Erhardt, L. Ferretti, N. Flavin, Q. Gao, M. Georges, R. Gurung, B. Harlizius, G. Hawkins, J. Hetzel, T. Hirano, D. Hulme, C. Jorgensen, M. Kessler, B. W. Kirkpatrick, B. Konfortov, S. Kostia, C. Kuhn, J. A. Lenstra, H. Leveziel, H. A. Lewin, B. Leyhe, L. Lil, I. Martin Burriel, McGraw, J. R. Miller, D. E. Moody, S. S. Moore, S. Nakane, I. J. Nijman, I. Olsaker, D. Pomp, A. Rando, M. Ron, A. Shalom, A. J. Teale, U. Thieven, B. G. D. Urquhart, D.-I. Vage, A. Van de Weghe, S. Varvio, R. Velmala, J. Vilkki, R. Weikard, C. Woodside, J. E. Womack, M. Zanotti and Zaragoza. 1997. A medium-density genetic linkage map of the bovine genome. Mamm. Genome 8:21-28.

Blouin, M. S. 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations, Trends Ecol. Evol. 18:503-511.

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.

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

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

Gasbarra, D., M. J. Sillanpaa and E. Arjas. 2005 Backward simulation of ancestors of sampled individuals. Theor. Popul. Biol. 67:75-83.

Guttorp, P. 1995. Stochastic modeling of scientific data. Chapman

and Hall, CRC press.
Hayes, B. J. and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33:209-229.
Hayes, B. J. and M. E. Goddard. 2008. Technical note: Prediction of breeding values using marker-derived relationship matrices. J. Anim. Sci. 86:2089-2092.
Hill, W. G. and B. S. Weir. 2007. Prediction of multi-locus inbreeding coeffcients and relation to linkage disequilibrium in random mating populations. Theor. Popul. Biol. 72:179-185.
Libiger, O. and N. J. Schork. 2007. A simulation-based analysis of chromosome segment sharing among a group of arbitrarily related individuals. Eur. J. Hum. Genet. 15:1260-1268.
Liu, B-H. 1998. Statisitical genomics. CRC press.
Lynch, M. and B. Walch. 1998. Genetics and analysis of quantitative traits. Sinauer Associates Inc, Sunderland, MA.
Macleod, I. M., B. J. Hayes and M. E. Goddard. 2006 Efficiency of dense bovine single-nucleotide polymorphisms to detect and position quantitative trait loci. Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brazil, August 13-18, 2006. CD-ROM communication no. 20-04.
Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, T. S. Sonstegard, T. P. L. Smith, S. S. Moore and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One. (submitted).
Meuwissen, T. H. E., B. Hayes and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.
Peng, B., C. I. Amos and M. Kimmel. 2007. Forward-time simulations of human populations with complex diseases. PLoS Genetics 3:e47.
Peng, B. and M. Kimmel. 2005. SimuPOP: a forward-time population genetics simulation environment. Bioinformatics 21:3686-3687.
Quass, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32:949-953.
Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218-223.
Solberg, T. R., A. K. Sonesson, J. A. Woolliams and T. H. E. Meuwissen. 2008. Genomic selection using different marker types and densities. J. Anim. Sci. 86:2447-2454.
Strand, A. E. 2002. METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics. Mol. Ecol. Notes 2:373-376.
Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard and P. M. Visscher. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17:520-526.
TeMeerman, G. J. and M. A. Van der Meulen. 1997. Genomic sharing surrounding alleles identical by descent: effects of genetic drift and population growth. Genet. Epidemiol. 14: 1125-1130.
VanRaden, P. M. 2007. Genomic measures of relationship and inbreeding. INTERBULL bulletin. 37: 33-36
VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.
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 prediction for north american Holstein bulls. J. Dairy Sci. 92:16-24.
Van Tassell, C. P., T. P. L. Smith, L. K. Matukumallik, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5:247-252.
Villanueva, B., R. Pong-Wong, J. Fernandez and M. A. Toro. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. J. Anim. Sci. 83:1747-1752.
Visscher, P. M., S. E. Medland, M. A. Ferreira, K. I. Morley, G. Zhu, B. K. Cornes, G. W. Montgomery and N. G. Martin. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2:e4.
Wu, R., C-X Ma and G. Casella. 2007. Statistical genetics of quantitative traits. Springer.