Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle

M. Pszczola,*†‡¹ H. A. Mulder,† and M. P. L. Calus†

*Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, the Netherlands

†Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, the Netherlands

‡Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Wolynska 33, 60-637 Poznan, Poland

ABSTRACT

Genomic selection (GS) permits accurate breeding values to be obtained for young animals, shortening the generation interval and accelerating the genetic gain, thereby leading to reduced costs for proven bulls. Genotyping a large number of animals using high-density single nucleotide polymorphism marker arrays is nevertheless expensive, and therefore, a method to reduce the costs of GS is desired. The aim of this study was to investigate an influence of enlarging the reference population, with either genotyped animals or individuals with predicted genotypes, on the accuracy of genomic estimated breeding values. A dairy cattle population was simulated in which proven bulls with 100 daughters were used as a reference population for GS. Phenotypic records were simulated for bulls with heritability equal to the reliability of daughter yield deviations based on 100 daughters. The simulated traits represented heritabilities at the level of individual daughter performance of 0.3, 0.05, and 0.01. Three scenarios were considered in which (1) the reference population consisted of 1,000 genotyped animals, (2) 1,000 ungenotyped animals were added to the reference population, and (3) the 1,000 animals added in scenario 2 were genotyped in addition to the 1,000 animals from scenario 1. Genotypes for ungenotyped animals were predicted with an average accuracy of 0.58. Additionally, an adjustment of the diagonal elements of the **G** matrix was proposed for animals with predicted genotypes. The accuracy of genomic estimated breeding values for juvenile animals was the highest for the scenario with 2,000 genotyped animals, being 0.90, 0.79, and 0.60 for the heritabilities of 0.3, 0.05, and 0.01, respectively. Accuracies did not differ significantly between the scenario with 1,000 genotyped animals only and the scenario in which 1,000 ungenotyped animals were added and the adjustment of the **G** matrix was applied. The absence of significant

increase in the accuracy of genomic estimated breeding values was attributed to the low accuracy of predicted genotypes. Although the differences were not significant, the difference between scenario 1 and 2 increased with decreasing heritability. Without the adjustment of the diagonal elements of the **G** matrix, accuracy decreased. Results suggest that inclusion of ungenotyped animals is only expected to enhance the accuracy of GS when the unknown genotypes can be predicted with high accuracy.

Key words: genomic selection, accuracy of genomic breeding value, prediction of genotypes, dairy cattle

INTRODUCTION

Conventionally, breeding value estimation for dairy cattle uses phenotypic performance of an animal itself and its relatives. Estimated breeding values are generally obtained using the BLUP method (Henderson, 1984). Achieving high accuracy of EBV for bulls requires collection of phenotypic data of the offspring. An important objective in dairy cattle breeding is therefore to shorten the generation interval while maintaining the accuracy of EBV at an acceptable level. Genomic selection (**GS**) as proposed by Meuwissen et al. (2001) is today adopted in many dairy cattle breeding programs to achieve this goal.

Linkage disequilibrium (**LD**) between adjacent loci exists because of phenomena such as selection, drift, and admixture, (Falconer and Mackay, 1996). Animals can now be genotyped with large SNP marker arrays, which cover the whole genome with high marker density. Because of the high marker density, it is expected that several SNP markers are in LD with QTL (Meuwissen et al., 2001). Quantitative trait loci effects can therefore be associated with single SNP markers or a group of markers. Consequently, given the SNP genotype of an animal, their phenotype can be predicted. Combining genotypic, phenotypic, and possibly pedigree information gives an opportunity to obtain an accurate EBV of an animal at an early stage in life, which can be used to perform GS, and thus, shorten the generation

Received October 20, 2009.

Accepted September 21, 2010.

¹Corresponding author: mbee@jay.up.poznan.pl

Table 1. Simulation parameters

Parameter	Value		
Effective population size of the first 1,000 generations	400		
Simulated genome length, M	6		
Number of simulated chromosomes	12		
Length of simulated chromosomes, M	0.5		
Distance between adjacent markers in generation 1,008, M	0.0012		
Number of SNP markers per cM	~7.5		
Number of QTL per cM	~0.33		
Minor allele frequency	0.29		
	Trait 1	Trait 2	Trait 3
Heritability of phenotype	0.3	0.05	0.01
Heritability for daughter yield deviation	0.89	0.56	0.20

interval. This can greatly accelerate the genetic gain and considerably reduce the costs of obtaining reliable breeding values for bulls (Meuwissen et al., 2001; Schaeffer, 2006).

Different approaches of including genomic information into breeding value estimations have been presented (Meuwissen et al., 2001; Kolbehdari et al., 2007; Long et al., 2007; Muir, 2007). One of them is the genomic BLUP procedure (**G-BLUP**), in which equal variances are assumed for all SNP effects (Meuwissen et al., 2001). The accuracy of the obtained genomic estimated breeding values (**GEV**) for juvenile animals that have no phenotypic observations is higher than the accuracy of pedigree indexes (Meuwissen et al., 2001). In the G-BLUP procedure, the genomic relationship matrix (**G**; VanRaden, 2008) can be used. Modeling SNP with equal variance is equivalent to using a genomic relationship matrix (VanRaden, 2008; Goddard, 2009; Strandén and Garrick, 2009). The **G** matrix contains relationship coefficients among evaluated animals estimated based on genomic information. Genomic relationship coefficients are estimated with higher accuracy than when using pedigree information, because genomic information allows the capture of Mendelian sampling across the genome.

The accuracy of GEBV for juvenile animals increases with the number of animals included in the reference population used to estimate SNP effects (Meuwissen et al., 2001; Goddard, 2009; Hayes et al., 2009). Enlarging the reference population, however, implies the genotyping of additional animals, which increases the costs of GS. Enlarging the reference population with ungenotyped animals for which genotypes can be predicted might be an inexpensive method to increase the accuracy of GEBV. Gengler et al. (2007) proposed a method that treats (unknown) genotypes as (missing) phenotypes and uses the additive relationship matrix, based on pedigree information, to predict genotypes. Although this is a promising and easy to implement strategy, the eventual influence of using predicted geno-

types in the reference population on the accuracy of GEBV has not yet been studied.

The objective of this study was to investigate the effect on the accuracy of GEBV of enlarging the reference population in a dairy cattle breeding program by adding bulls with known or predicted genotypes.

MATERIALS AND METHODS

Simulation

A dairy cattle population was simulated using similar assumptions as in Villumsen et al. (2009) and Calus et al. (2008). Data sets were simulated for daughter yield deviations (**DYD**) based on observations of 100 daughters of traits with high (0.3), moderate (0.05), and low (0.01) heritability. The first 1,000 generations had an effective population size of 400, consisting of 200 sires and 200 dams. All loci had alleles 1 and 2 segregating in the first generation, both with an allele frequency of 0.5. Linkage disequilibrium was established by performing random mating for the first 1,000 generations. Inheritance of parental chromosomes was in accordance with Haldane's (1919) mapping function. The mutation rate was 2×10^{-5} , where a mutation in allele 1 (2) yielded an allele 2 (1).

Generated genome length was 6 M and the genome consisted of 12 chromosomes, each 0.5 M long. This corresponds, approximately, to 20% of the actual cattle genome (Ihara et al., 2004). Marker loci (5,002) were spaced at fixed distances of 0.12 cM across the genome. After 1,000 generations of random mating, on average 4,500 markers were still segregating (i.e., on average 7.5 SNP/cM). Between 198 and 208 SNP (1 every 20 SNP) were removed from the segregating loci and used as QTL. Parameters used in the simulation are summarized in Table 1.

After the first 1,000 generations (i.e., in generation 1,001), the population was extended to 800 individuals. In generations 1,001 to 1,008, no mutations were

Figure 1. Simulation scheme.

simulated. In generations 1,001 to 1,007, 50 males and 200 females were randomly chosen as parents of the next generation. The matings were restricted such that there were no full-sibs among the 800 offspring. For generations 1,002 to 1,008, genotypes, true breeding values (**TBV**), and phenotypes of the males were simulated. Generation 1,008, containing juvenile animals, was simulated with unknown phenotypes. Pedigree was stored for all animals from generations 1,002 to 1,008. The outline of the simulation is presented in Figure 1.

To simulate TBV, QTL effects were randomly drawn from a normal distribution, $N(0, 1)$. All QTL effects were summed per animal to obtain TBV, assuming that QTL were independent. The total variance of the TBV was calculated (denoted as σ_{TBV}^2) as a variance of TBV across all animals. Phenotypes were obtained by adding a random residual term, $N(0, \sigma_e^2)$, to the TBV; σ_e^2 was derived as

$$\sigma_e^2 = \frac{\sigma_{TBV}^2(1 - h^2)}{h^2}, \quad [1]$$

where h^2 is the simulated heritability (see Table 1). Phenotypic records were simulated for bulls, with an h^2 equal to the reliability of DYD. This reliability (r_{IH}^2) for progeny-tested bulls was calculated according to the formula of Mrode (2005):

$$r_{IH}^2 = \frac{\frac{1}{4}nh^2}{1 + \frac{1}{4}(n-1)h^2}, \quad [2]$$

where n is the number of daughters, which was considered to be 100. As a result, heritabilities of 0.3, 0.05, and 0.01 at the phenotypic level yield heritabilities of 0.89, 0.56, and 0.20 at the DYD level, respectively (Table 1).

Scenarios

To meet the goal of this study the described simulation was performed for all 3 heritability levels, and each simulation was replicated 10 times. Three scenarios were considered with, on the one hand, different sizes of the reference population and, on the other hand, different numbers of animals with known or predicted genotypes. In the first scenario, GEBV were estimated using a reference population consisting of 1,000 genotyped sires chosen randomly from genotyped bulls within generations 1,002 to 1,007 (200 sires out of 400 available genotyped bulls) with phenotypes based on DYD.

In the second scenario, GEBV were estimated with an additional 1,000 ungenotyped bulls. Those bulls were the remaining 200 bulls per generation after choosing bulls for scenario 1 and had no offspring. A particular genotyped bull was mated only once with a particular cow, resulting in one offspring; the highest possible degree of relationship among the ungenotyped and genotyped animals was therefore half-sib or parent-offspring. The unknown genotypes were predicted using the method of regression on gene content, proposed by Gengler et al. (2007). This method treats (missing) genotypes as (unknown) phenotypes and uses the additive genetic relationship matrix (**A**) to predict them using the following model for each SNP separately:

$$gc_i = u_{gc} + d_i + e_{gc_i}, \quad [3]$$

where gc is the observed (missing) gene content for (un) genotyped animal i , μ_{gc} is an overall mean, d is EBV for gene content, and e_{gc} is the residual of gene content. The **A** matrix contained animals from generations 1,002 to 1,008. The heritability used in the mixed model equations was 0.99. The ASREML software (Gilmour et al., 2002) was used to solve the mixed model equations.

The third scenario was similar to the second one, except that all 2,000 bulls were considered genotyped. A traditional BLUP (scenario 4) was also performed using phenotypes for the same 2,000 bulls considered in scenario 3.

Estimation of GEBV

After each simulation, genotypes of the animals were used to create the $\mathbf{G}$ matrix according to VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2 \sum p_i(1 - p_i)}, \quad [4]$$

where p_i is the frequency of the second allele at locus i , and $\mathbf{Z}$ is derived by subtracting 2 times the allele frequency expressed as a difference of 0.5; that is, $2(p_i - 0.5)$, from matrix $\mathbf{M}$ that specifies the marker genotypes for each individual as -1 , 0 , or 1 (VanRaden, 2008). In this study the allele frequencies p_i were considered to be 0.5; therefore, $\mathbf{Z}$ was the same as $\mathbf{M}$. The value for p_i of 0.5 reflects the allele frequency in our simulated base population. In preliminary analysis, we found limited differences when using 0.5 instead of the allele frequency in the current population for p_i . VanRaden (2008) found that using different values for p_i hardly affects the accuracy of the GEBV. However, Aguilar et al. (2010) and Christensen and Lund (2010) found that using different allele frequencies does influence the accuracy of GEBV. Results of Aguilar et al. (2010) indicated that allele frequencies of 0.5 actually gave the highest accuracy. These different results indicate that the effect of used allele frequencies for p_i appears to be different for different data sets.

Subsequently, GEBV were estimated with G-BLUP using the following model:

$$y_i = \mu + a_i + e_i, \quad [5]$$

where μ is an overall mean, a_i is an EBV, and e_i is the random error term. The EBV were assumed to be distributed as $N(0, \mathbf{G}\sigma_a^2)$, and the residuals were assumed to be distributed as $N(0, \sigma_e^2)$. The genetic variance σ_a^2 and residual variance σ_e^2 were estimated using REML implemented in ASReml software (Gilmour et al., 2002).

In G-BLUP, an inverse of the $\mathbf{G}$ matrix is required. To avoid singularities in $\mathbf{G}$, the latter was weighted by $\mathbf{A}$ as follows: $\mathbf{G}\omega = \omega\mathbf{G} + (1 - \omega)\mathbf{A}$ (VanRaden, 2008), using a weighting factor (ω) of 0.99, meaning that a relatively low weight was given to the $\mathbf{A}$ matrix. The $\mathbf{A}$ matrix used to weight the $\mathbf{G}$ matrix contained only animals that were present in $\mathbf{G}$.

Comparison of GEBV

The 10 replicates were analyzed by calculating the accuracies of GEBV, regression of simulated breeding

values on GEBV, and mean squared errors of prediction (**MSEP**) for each group of animals. Furthermore, estimated heritabilities were compared with the simulated values.

Adjustment of Diagonal Elements of the G Matrix

Initial analyses (Figure 2) showed that the diagonal elements of the $\mathbf{G}$ matrix for ungenotyped animals using predicted gene contents were much lower than 1.0. This is in conflict with the expectation of diagonal elements of a relationship matrix, which is 1 plus the inbreeding coefficient of the animal (Wright, 1922). Although the expectation of the diagonal elements of the $\mathbf{G}$ matrix is >1 , the diagonal elements in the $\mathbf{G}$ matrix can be lower for some animals (VanRaden, 2008). However, in our case, the average diagonal element of $\mathbf{G}$ for ungenotyped animals was only 0.73. The low values for the diagonal elements of the $\mathbf{G}$ matrix were caused by the fact that the predicted gene content is regressed back to the mean for ungenotyped animals. This regression to the mean is due to the nature of BLUP and leads to an excess of expected heterozygotes. Both off-diagonal and diagonal elements were affected by the regression to the mean; nevertheless, in absolute terms the effect was largest on diagonal elements. The mean absolute difference of predicted and true values of the off-diagonals was only 0.03. Because the diagonal elements of $\mathbf{G}$, based on the number of homozygous loci of an animal (VanRaden, 2008), were too low, we proposed an adjustment of the diagonal elements for ungenotyped animals. In this adjustment, the diagonal elements for ungenotyped animals are calculated, assuming no other relationships between sire and dam except between sire and maternal grandsire, as

$$g_{jj} = 1 + 0.25 \times g_{(s\ j, mgs\ j)}, \quad [6]$$

where g_{jj} is the diagonal element for the ungenotyped animal j , and $g_{(s\ j, mgs\ j)}$ is the genomic relationship coefficient between the genotyped sire (s) and genotyped maternal grandsire (mgs) of animal j . Results obtained after this adjustment were denoted as an additional scenario (scenario 2a), whereas the results obtained without the adjustment were presented as scenario 2.

RESULTS

Characteristics of the Simulation

In each replicate, animals with genotypes were simulated. The average distance between adjacent loci across the whole genome, calculated in generation 1,008, was 0.13 cM (Table 1). The average LD between adjacent

Table 2. Estimated heritabilities of daughter yield deviation across different scenarios and simulated for both heritabilities

Scenario ¹	High simulated heritability		Moderate simulated heritability		Low simulated heritability	
	h^2	SE ²	h^2	SE	h^2	SE
Simulated heritability	0.89	—	0.56	—	0.20	—
1	0.90 ^a	0.04	0.58 ^c	0.01	0.22 ^e	0.05
2	0.63 ^b	0.03	0.48 ^d	0.03	0.23 ^e	0.04
2a	0.92 ^a	0.04	0.60 ^c	0.01	0.24 ^e	0.04
3	0.91 ^a	0.03	0.61 ^c	0.03	0.25 ^e	0.03
4	0.92 ^a	0.03	0.57 ^c	0.03	0.21 ^e	0.04

^{a-c}Values with identical superscripts did not differ significantly among scenarios ($P > 0.05$); standard errors of 10 replicates ranged from 0 to 0.012.

¹Scenario 1 consisted of 1,000 genotyped animals; scenario 2 consisted of 1,000 genotyped and 1,000 ungenotyped animals with unadjusted diagonal elements of the **G** matrix for ungenotyped animals; scenario 2a consisted of 1,000 genotyped and 1,000 ungenotyped animals with adjusted diagonal elements of the **G** matrix for ungenotyped animals; scenario 3 consisted of 2,000 genotyped animals; scenario 4 consisted of 2,000 genotyped animals analyzed with use of traditional BLUP.

²Standard error of estimated heritability averaged over 10 replicates.

markers measured as r^2 (Hill and Robertson, 1968) was 0.41. Minor allele frequency, averaged across all marker loci in generation 1,008, was 0.29. Gene contents for 1,000 animals were predicted with an average accuracy of 0.58.

Estimated Heritabilities

Estimated DYD heritabilities averaged over 10 replicates for scenarios 1, 2, 3, and 4 (Table 2) were always slightly higher than the simulated values. Differences among the estimated DYD heritabilities for the differ-

ent scenarios, apart from scenario 2 for high and moderate heritability, were not statistically significant from the simulated value. Estimated heritability in scenario 2 for high and moderate heritability traits was significantly lower from the simulated value. The adjustment of the diagonal elements of the **G** matrix in scenario 2a resulted in estimated heritabilities not significantly different from the simulated values.

Evaluation of **G** Matrix

The diagonal elements of the **G** matrix, without adjusting the coefficients for predicted genotypes, were considerably lower when compared with the same coefficients for genotyped animals (Figure 2). The adjustment proposed in the present study resulted in diagonal elements that were closer to their expectations, as can be seen from Figure 3. Those coefficients were, nevertheless, still lower than for genotyped animals; the coefficients for the group of the additional animals were generally lower in the **G** matrix based on predicted gene contents than the corresponding coefficients when genotype data were available (Figure 3). The coefficients in the **G** matrix were higher than pedigree-based coefficients, for genotyped as well as the additional 1,000 bulls with predicted genotypes (Figures 4 and 5).

Evaluation of GEBV

The accuracy of GEBV for the first 1,000 animals for all simulated heritability levels was the highest in scenario 3 (0.96 for high, 0.88 for moderate, and 0.72 for low heritability; Table 3). The lowest values were obtained in scenario 2 for high heritability (0.93) and for scenario 4 for moderate (0.81) and low (0.60) heritability. The accuracy of GEBV in scenarios 1 and 2a did not differ significantly from each other ($P > 0.05$)

Figure 2. Coefficients of the genomic relationship matrix for animals with predicted genotypes plotted against the same coefficients calculated using their true genotypes without adjusting the diagonal elements.

Figure 3. Coefficients of the genomic relationship matrix for animals with predicted genotypes plotted against the same coefficients calculated using their true genotypes, after adjusting diagonal elements of **G** matrix for animals with predicted genotypes.

within high and moderate heritability and was significantly different for low heritability ($P < 0.05$).

The highest accuracies for the additional group of the animals were observed in scenario 3 for high (0.96), moderate (0.87), and low (0.70) heritabilities; these estimates were significantly different from the accuracies in scenario 2, 2a, and 4 ($P < 0.05$). For all heritability levels, accuracies of scenarios 2 and 4 were significantly different from each other, whereas the accuracy in scenarios 2a and 4 were not significantly different ($P > 0.05$).

Adjusting the diagonal elements of the **G** matrix (scenario 2a) resulted in a significant increase of the accuracy of GEBV for the additional group of animals for low and moderate heritability, and for all groups of animals for high heritability when compared with the results of the scenario with no adjustment (scenario 2). For low heritability, a significant decrease was observed in the group of the first 1,000 animals.

Similarly to the group of first 1,000 animals and the group of additional animals, the accuracy for the juvenile animals was the highest in scenario 3 for all 3 heritability levels: 0.90 (high), 0.79 (moderate) and 0.60 (low), and the lowest in scenario 4: 0.57 (high), 0.48 (moderate), and 0.33 (low). Accuracies reached in scenarios 1 and 2a were somewhat lower than those in scenario 3. Accuracy estimates for the first 1,000 animals in scenario 2 were significantly higher ($P < 0.05$) than those in scenario 1 for low heritability, lower

for high heritability, and not significantly different from each other ($P > 0.05$) for moderate heritability. The accuracy of juvenile animal GEBV in scenario 2 was significantly lower ($P < 0.05$) than in scenario 1 for high and not significantly different ($P > 0.05$) for the other 2 heritability levels. The adjustment of the diagonal elements of the **G** matrix (scenario 2a) resulted in significantly higher ($P < 0.05$) accuracy for all groups of the animals for high and low heritability and for the group of the 1,000 additional ungenotyped animals for moderate heritability when compared with scenario 2. The increase for the first 1,000 animals and juveniles for moderate heritability was, however, not significant ($P > 0.05$). Differences in accuracies between scenarios 1 and 2a were not significant across all heritability levels ($P > 0.05$), although the difference in accuracy between scenarios 2a and 1 tended to increase with decreasing heritability.

In summary, the accuracies of GEBV were the highest when the true genotypes of the additional 1,000 bulls were added to the reference population and were considerably lower for traditional BLUP. The accuracies of the GEBV for the animals with predicted genotypes were lower than for the genotyped ones and similar to the accuracies when applying traditional BLUP. Nevertheless, after adding animals with predicted genotypes and adjusting the diagonal coefficients of the **G** matrix, the accuracy of GEBV did not decline compared with the scenario when no animals were added. Furthermore,

Figure 4. Coefficients of the genomic relationship matrix for the animals with predicted genotypes plotted against pedigree-based relationship coefficients, with adjusted diagonal elements of **G** matrix for animals with predicted genotypes applied.

Table 3. Accuracies (Acc.), regression coefficients (Reg.), and mean squared error of prediction (MSEP) of genomic EBV for groups of 1,000 first, additional, and juvenile animals for heritability of 0.3 (0.89 daughter yield deviation) for all scenarios

Scenario ¹	First 1,000 animals			Additional animals			Juvenile animals		
	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP
1	0.96 ^a	1.02	6.79	—	—	—	0.84 ^h	1.01	23.09
2	0.93 ^b	1.15 ²	12.35	0.65 ^e	1.27 ²	45.15	0.80 ⁱ	1.19 ²	28.47
2a	0.96 ^a	1.00	6.50	0.95 ^f	1.03	8.50	0.83 ^h	0.98	22.63
3	0.96 ^c	1.02	5.54	0.96 ^g	1.02	6.07	0.90 ^j	1.01	15.09
4	0.95 ^d	1.01	7.79	0.95 ^f	1.00	8.18	0.57 ^k	1.04	52.25

^{a-k}Values with identical superscripts did not differ significantly ($P > 0.05$); standard errors of 10 replicates ranged from 0 to 0.02 for Acc. and Reg., and from 0.2 to 1.77 for MSEP.

¹Scenario 1 consisted of 1,000 genotyped animals; scenario 2 consisted of 1,000 genotyped and 1,000 ungenotyped animals with unadjusted diagonal elements of the **G** matrix for ungenotyped animals; scenario 2a consisted of 1,000 genotyped and 1,000 ungenotyped animals with adjusted diagonal elements of the **G** matrix for ungenotyped animals; scenario 3 consisted of 2,000 genotyped animals; scenario 4 consisted of 2,000 genotyped animals analyzed with use of traditional BLUP.

²Reg. significantly different from 1.

the accuracies of juvenile animals increased in scenario 2a compared with scenario 1 (Tables 4 and 5) for traits with moderate and low heritability, although the difference was not significantly greater than zero.

For all simulated heritability levels and across all scenarios, regression coefficients of true on estimated breeding values were close to 1 (Tables 3 to 5). For the juvenile animals, regression coefficients were generally slightly less than 1. Regression coefficients of the animals with predicted genotypes were in all cases greater than 1, indicating that the variance of their GEBV was underestimated. This is caused by the fact that estimat-

ed gene contents were shrunk back toward the mean. Regression coefficients were significantly greater than 1 in scenario 2, especially for the group of the additional animals. These differences from 1 observed in scenario 2, apart from the group of additional animals for moderate heritability, were no longer significant when the diagonal elements were adjusted (scenario 2a); MSEP ranged from 5.54 to 68.88 across all scenarios.

DISCUSSION

The objective of this study was to investigate the effect of enlarging the reference population in a dairy cattle breeding program, by adding bulls with known or predicted genotypes, on the accuracy of GEBV. Four scenarios were evaluated that differed with regard to the number of genotyped animals and ungenotyped animals with predicted genotypes. As expected, the accuracies of GEBV for all traits were higher for the scenario with a higher number of genotyped animals. A similar trend was observed by Goddard (2009).

Accuracy of GEBV

In general, adding animals with predicted genotypes to the reference population did not significantly increase the accuracy of GEBV; however, a trend of increasing difference in accuracy between scenarios 1 and 2a with decreasing heritability was observed (Tables 3, 4, and 5). This trend suggests that when using traits with low heritability or less accurate phenotypic records (i.e., own performance records), scenario 2a, in which animals with predicted genotypes were added, may become beneficial.

Comparison of the estimates for BLUP (scenario 4) and G-BLUP in scenario 2a showed that, in general, accuracy increased significantly for the group of the first

Figure 5. Coefficients of the genomic relationship matrix for the genotyped animals plotted against pedigree-based relationship coefficients.

Table 4. Accuracies (Acc.), regression coefficients (Reg.), and mean squared prediction errors (MSEP) of genomic EBV for groups of 1,000 first, additional, and juvenile animals for heritability of 0.05 (0.56 daughter yield deviation) for all scenarios

Scenario ¹	First 1,000 animals			Additional animals			Juvenile animals		
	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP
1	0.85 ^a	1.02	22.76	—	—	—	0.70 ^g	0.99	40.06
2	0.85 ^a	1.07 ²	22.56	0.62 ^d	1.23 ²	49.67	0.70 ^g	1.06	40.03
2a	0.85 ^a	1.01	20.93	0.79 ^f	1.12 ²	29.65	0.72 ^g	0.96	37.79
3	0.88 ^b	1.00	18.01	0.87 ^e	1.00	19.20	0.79 ^h	0.97	30.33
4	0.81 ^c	0.99	27.75	0.79 ^f	0.99	29.52	0.48 ⁱ	0.99	60.95

^{a-i}Values with identical superscripts did not differ significantly ($P > 0.05$); standard errors of 10 replicates ranged from 0 to 0.03 for Acc. and Reg., and from 0.88 to 2.41 for MSEP.

¹Scenario 1 consisted of 1,000 genotyped animals; scenario 2 consisted of 1,000 genotyped and 1,000 ungenotyped animals with unadjusted diagonal elements of the **G** matrix for ungenotyped animals; scenario 2a consisted of 1,000 genotyped and 1,000 ungenotyped animals with adjusted diagonal elements of the **G** matrix for ungenotyped animals; scenario 3 consisted of 2,000 genotyped animals; scenario 4 consisted of 2,000 genotyped animals analyzed with use of traditional BLUP.

²Reg. significantly different from 1.

1,000 animals and did not differ for the group of additional ungenotyped animals. For the juvenile animals, however, superiority of G-BLUP can be seen clearly, as also reported by others (Meuwissen et al., 2001; Schaeffer, 2006). The improvement in accuracy for juvenile animals comparing G-BLUP with traditional BLUP was apparent and similar at all heritability levels. This is in contrast to findings of others, who found that use of marker information was especially beneficial for low heritability traits (Meuwissen and Goddard, 1996; Meuwissen et al., 2001; Mulder et al., 2010).

The coefficients of the **G** matrix between animals with predicted gene contents were clearly biased as they were shrunk toward the mean. A solution could be to account for the variance of predicted gene content. The bias of the coefficients can be judged based on Figure 2 as the difference between the values on the y- and x-axes. Although calculation of the diagonal elements in the **G** matrix using equation [5] increased these coefficients on average by 0.45, Figure 3 shows

that the variance of those adjusted coefficients is still underestimated. The adjustment of the diagonals, however, did result in a clear improvement of the accuracy of the GEBV for scenario 2a compared with scenario 2. This implies that, although parts of the **G** matrix may still be biased, our ad hoc adjustment has made the **G** matrix overall more consistent, which resulted in more accurate prediction of the GEBV. Therefore, in situations where parts of the **G** matrix are dependent on different sources of information with different levels of accuracy, such as known versus predicted gene content, it appears to be important to ensure that different parts of the matrix have similar properties. Christensen and Lund (2010) proposed a model in which, unlike the method presented here, both predicted genotypes and the variance of the prediction are included. This strategy, therefore, directly yields a more consistent matrix.

When the aim is to increase the accuracy of GEBV for a certain group of juveniles, the best strategy is

Table 5. Accuracies (Acc.), regression coefficients (Reg.), and mean squared prediction errors (MSEP) of genomic EBV for groups of 1,000 first, additional, and juvenile animals for heritability of 0.01 (0.2 daughter yield deviation) for all scenarios

Scenario ¹	First 1,000 animals			Additional animals			Juvenile animals		
	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP	Acc.	Reg.	MSEP
1	0.65 ^a	1.03	45.72	—	—	—	0.52 ^j	1.01	56.14
2	0.69 ^b	1.02	42.04	0.51 ^f	1.18 ²	58.19	0.54 ⁱ	0.98	54.94
2a	0.66 ^c	0.95	42.02	0.57 ^g	1.08	53.04	0.56 ^j	0.96	55.03
3	0.72 ^d	0.99	38.85	0.70 ^h	0.99	40.35	0.60 ^j	0.94	49.57
4	0.60 ^e	1.00	50.78	0.57 ^g	1.00	53.53	0.33 ^k	0.91 ²	68.88

^{a-k}Values with identical superscripts did not differ significantly ($P > 0.05$); standard errors of 10 replicates ranged from 0 to 0.07 for Acc. and Reg., and from 1.93 to 2.71 for MSEP.

¹Scenario 1 consisted of 1,000 genotyped animals; scenario 2 consisted of 1,000 genotyped and 1,000 ungenotyped animals with unadjusted diagonal elements of the **G** matrix for ungenotyped animals; scenario 2a consisted of 1,000 genotyped and 1,000 ungenotyped animals with adjusted diagonal elements of the **G** matrix for ungenotyped animals; scenario 3 consisted of 2,000 genotyped animals; scenario 4 consisted of 2,000 genotyped animals analyzed with use of traditional BLUP.

²Reg. significantly different from 1.

probably to add animals to the reference population that are closely related to those juveniles (e.g., their parents; Habier et al., 2010). When the aim is to increase the accuracy of GEBV for juveniles throughout the population, one strategy may be to add animals that are not closely related to the reference population, and therefore add to the average relationship of the reference population to any given animal in the population. However, when animals with predicted genotypes are added to the reference population, the gain may be larger when those animals are more closely related to the reference population, because their genotypes are predicted with higher accuracy.

Figures 4 and 5 show that coefficients of the **A** matrix were lower than those of the **G** matrix, which is because of the difference in the level of inbreeding when using pedigree or genomic information. The level of inbreeding when using pedigree data was calculated using the first generation in the pedigree as the base generation, while the genomic inbreeding level is calculated using generation 1 in the simulation as base generation. Therefore, inbreeding coefficients in the **G** matrix are higher, and consequently the coefficients in the **G** matrix are higher compared with the corresponding ones in the **A** matrix.

Accuracy of Genomic Relationship Matrix with Predicted Gene Contents

In the present study, ungenotyped animals did not have offspring and therefore no genotype information on descendants was available. A particular bull was mated only once with a particular cow and, therefore, the highest possible degree of relationship among the ungenotyped and genotyped animals was half-sib or parent-offspring. This resulted in relatively low accuracy of predicted gene content and, thus, low accuracy of GEBV for ungenotyped and juvenile animals in scenario 2. To increase the accuracy of the prediction of genotypes, ungenotyped animals should be chosen that have genotyped offspring available. In such a scenario, ungenotyped animals could be, for instance, dams of genotyped offspring. The accuracy of predicted gene content is then expected to be 0.88 with 10 genotyped half-sib offspring, using the square root of equation [2] and assuming a heritability of 1.0 for gene content. In that case, superior accuracies of GEBV to the ones obtained in this study may be obtained. A disadvantage is that the phenotypic information of dams generally has a lower reliability compared with bulls.

Alternatively, if available, the additional genotypic information could be used for prediction of genotypes. The use of this additional information could lead to the maximum possible accuracy of 0.707 when both parents

but no offspring are genotyped, and assuming that the heritability is 1.0 ($r = \sqrt{\mathbf{b}\mathbf{G}}/\sigma_a = \sqrt{\frac{1}{2}}$, where both **b** and **G** contain values of 1/2 and $\sigma_a = 1$; Mrode, 2005). With genotyped offspring, this accuracy could be substantially higher; Gengler et al. (2008), for instance, reported an accuracy of 0.93 when Canadian Holstein data were used to evaluate the gene content prediction method.

A similar strategy to include ungenotyped animals is to combine genotypic and pedigree information in a modified relationship matrix. Such an approach was proposed by Legarra et al. (2009) and Christensen and Lund (2010), who used genomic information to enrich the **A** matrix whenever this information was available. This approach does not require explicit prediction of the unknown genotypes, but directly predicts the genomic relationships using pedigree and genomic information from relatives, and therefore provides a one-step genetic evaluation with use of genomic information. This alternative approach may be able to overcome the problems encountered due to shrinkage of estimated gene content in the method applied here. An important advantage of this approach is that breeding values can be estimated in a single step including records of all genotyped and ungenotyped animals, in contrast to our 2-step approach in which DYD are used to predict GEBV. In the simulation presented by Christensen and Lund (2010), a higher accuracy of GEBV was obtained for the 1-step approach compared with the 2-step approach. This suggests that part of the information in the 2-step approach is lost due to errors in DYD prediction, thereby leading to lower accuracy of the GEBV.

Inversion of G Matrix

The G-BLUP implemented in a traditional mixed model equations procedure requires the inverted **G** matrix. When the number of animals in the data set is large, direct inversion of **G** may not be possible because of computational difficulties. Furthermore, possible singularities in the **G** matrix make unique inversion impossible. Singularities in the **G** matrix may appear among the animals with predicted genotypes when full-sibs are present or genotypes were predicted for the animals that do not have offspring and have unknown parents. In both cases, the method of Gengler et al. (2007) gives the same estimates for these groups of the animals. Therefore, no full-sibs were simulated among the ungenotyped animals in the present study. To deal with any further singularities, the final **G** matrix was a weighted sum of the **G** and **A** matrix, following VanRaden (2008). Another method to eliminate singularities from the matrix is to modify the diagonal of the

relationship matrix by adding a small number to it; for example, 10^{-6} (Zhong et al., 2009).

An alternative method that does not require the inversion of the relationship matrix was recently proposed by Misztal et al. (2009). In this method, the additive genetic relationship matrix is modified by a matrix accounting for genomic information, as in Legarra et al. (2009), and then used in an asymmetric set of mixed model equations (Harville, 1976; Henderson, 1984, 1985) solved with the algorithm proposed by Misztal et al. (2009). This method resolves the problem related to singularities in the relationship matrix by avoiding the inversion step.

Implications for Animal Breeding

This study confirmed that G-BLUP is more beneficial than traditional breeding value estimation. When ungenotyped animals are to be used to enlarge the reference population to increase accuracy of GEBV, animals should be chosen whose genotypes can be predicted with sufficiently high accuracy and have phenotypes of high quality. Fulfilling these conditions may lead to achieving high accuracy of predicted genotypes and make it worthwhile to enlarge the reference population with ungenotyped animals.

CONCLUSIONS

This study showed that inclusion of animals with predicted genotypes in the reference population did not significantly increase GEBV accuracies for juvenile animals. This lack of significance was mainly attributed to the low accuracy of predicted genotypes. Therefore, inclusion of ungenotyped animals is only expected to enhance the accuracy of GS when the unknown genotypes can be predicted with high accuracy.

ACKNOWLEDGMENTS

Marcin Pszczola gratefully acknowledges the financial support of the Koepon Foundation (Amersfoort, the Netherlands). Han Mulder was financially supported by CRV (Arnhem, the Netherlands), Hendrix Genetics (Boxmeer, the Netherlands), IPG (Beuningen, the Netherlands), and the European Commission, within the 6th Framework project SABRE, contract No. FOOD-CT-2006-016250. The text represents the authors' views and does not necessarily represent a position of the Commission, who will not be liable for the use made of such information. The authors gratefully acknowledge the meticulous work of two anonymous reviewers who significantly improved the manuscript.

REFERENCES

- Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93:743–752.
- 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.
- Christensen, O. F., and M. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42:2.
- Falconer, D. S., and T. F. C. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th ed. Longman, New York, NY.
- Gengler, N., S. Abras, C. Verkenne, S. Vanderick, M. Szydlowski, and R. Renaville. 2008. Accuracy of prediction of gene content in large animal populations and its use for candidate gene detection and genetic evaluation. *J. Dairy Sci.* 91:1652–1659.
- Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: Application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal* 1:21–28.
- Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham, and R. Thompson. 2002. *ASREML User Guide*, release 1.0. VSN International Ltd., Hemel Hempstead, UK.
- Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* 136:245–257.
- Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* 42:5.
- Haldane, J. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.* 8:299–309.
- Harville, D. 1976. Extension of the Gauss-Markov theorem to include the estimation of random effects. *Ann. Stat.* 4:384–395.
- 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.
- Henderson, C. R. 1984. *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Ontario, Canada.
- Henderson, C. R. 1985. Best linear unbiased prediction using relationship matrices derived from selected base populations. *J. Dairy Sci.* 68:443–448.
- Hill, W., and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226–231.
- Ihara, N., A. Takasuga, K. Mizoshita, H. Takeda, M. Sugimoto, Y. Mizoguchi, T. Hirano, T. Itoh, T. Watanabe, and K. Reed. 2004. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Res.* 14:1987–1998.
- Kolbehari, D., L. R. Schaeffer, and J. A. B. Robinson. 2007. Estimation of genome-wide haplotype effects in half-sib designs. *J. Anim. Breed. Genet.* 124:356–361.
- Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92:4656–4663.
- Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendaño. 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., and M. Goddard. 1996. The use of marker haplotypes in animal breeding schemes. *Genet. Sel. Evol.* 28:161–176.
- 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.
- Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. *J. Dairy Sci.* 92:4648–4655.
- Mrode, R. 2005. *Linear Models for the Prediction of Animal Breeding Values*. CABI, Wallingford, UK.
- Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under

- alternative trait and genomic parameters. *J. Anim. Breed. Genet.* 124:342–355.
- Mulder, H. A., T. H. E. Meuwissen, M. P. L. Calus, and R. F. Veerkamp. 2010. The effect of missing marker genotypes on the accuracy of gene-assisted breeding value estimation: A comparison of methods. *Animal* 4:9–19.
- Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.* 123:218–223.
- Strandén, I., and D. J. Garrick. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.* 92:2971–2975.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423.
- Villumsen, T. M., L. Janss, and M. S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126:3–13.
- Wright, S. 1922. Coefficients of inbreeding and relationship. *Am. Nat.* 56:330–338.
- Zhong, S., J. C. M. Dekkers, R. L. Fernando, and J.-L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics* 182:355–364.