The Accuracy Of Genomic Selection Using (Un)Genotyped Animals To Enlarge The Reference Population

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

Introduction

Genomic selection (GS) allows obtaining accurate breeding values for young animals, shortening generation interval and accelerating genetic gain, leading to reduced costs for proven bulls. Nevertheless, genotyping a large number of animals using a high density SNP chip is still expensive, and therefore, methods to reduce the costs of GS are desirable. Therefore, the aim of this study was to investigate the influence of enlarging the reference population, either by genotyped animals or individuals with predicted genotypes, on the accuracy of genomic estimated breeding values (GEBV).

Material and methods

Simulation. A dairy cattle population was simulated. Daughter yield deviations (DYD) were simulated for traits with high (0.3), moderate (0.05) and low (0.01) heritability. The first 1000 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. LD was established by performing random mating for the first 1000 generations. The mutation rate was 2×10^{-5}.
Generated genome length was 6 M and consisted of 12, equally long, chromosomes. 5002 marker loci were spaced at fixed distances of 0.12 cM across the genome. After 1,000 generations of random mating, on average 4500 markers were still segregating, i.e. ~7.5 SNP/cM. Between 198 and 208 SNPs were removed and used as QTL.
In generation 1001 the population was extended to 800 individuals. In generations 1001-1008 no mutations were simulated. In generations 1001-1007, 50 males and 200 females were randomly chosen as parents of the next generation, avoiding creating full-sibs. For generations 1002 -1008, genotypes, true breeding values (TBV), and phenotypes of the males were simulated. Generation 1008, containing juvenile animals, was simulated with unknown phenotypes. TBV were obtained by summing all QTL effects per animal. The QTL effects were drawn from a normal distribution. The variance of the true breeding values was calculated (denoted as σ^2_{TBV}). Phenotypes were obtained by adding a random residual term, $N(0,\sigma^2_e)$, to the TBV; σ^2_e was derived as σ^2_{TBV} multiplied by $(1-h^2)/h^2$, where h^2 is the simulated heritability. Instead of raw phenotypes, DYDs were simulated as a single phenotypic record. The heritability for those records, in order to account for the accuracy of

[*] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, The Netherlands
[†] Animal Breeding and Genomics Centre, Wageningen Institute of Animal Sciences Wageningen University, 6700 AH Wageningen, The Netherlands
[‡] Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Wolynska 33, 60-637 Poznan, Poland

DYD, was derived as reliability of selection (r^2_{IH}) for progeny-tested bulls, according to Mrode (2005). Each sire was assumed to have 100 daughters tested. 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. Simulations were replicated 10 times.

Scenarios. Three scenarios were considered with different sizes of the reference population and different numbers of the animals with known or predicted genotypes. In the first scenario GEBV were estimated using a reference population consisting of 1000 genotyped sires selected randomly from generations 1002-1007 (200 out of 400 sires per generation).

In the second scenario GEBV were estimated with an additional 1000 bulls (i.e. 200 per generation). The unknown genotypes were predicted using the regression on gene content method (Gengler et al. (2007)), where (missing) genotypes are treated as (unknown) phenotypes and are predicted using the additive genetic relationship matrix (**A**) for each SNP separately. The **A** matrix contained animals from generations 1002 to 1008. The heritability used was 0.99. ASReml (Gilmour et al. (2002)) was used to solve the mixed model equations.

The third scenario was similar to the second one; except that all 2000 used bulls were considered to be genotyped. A traditional BLUP (scenario 4) was performed using phenotypes for the same 2000 bulls considered in scenario 3.

Estimation of Genomic Estimated Breeding Values. The **G** matrix was created using the following formula by VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$$

Subsequently, GEBV were estimated using G-BLUP with use of the model comprising an overall mean, an estimated breeding value and the random error term. The estimated breeding values were assumed to be distributed as $N(0,\mathbf{G}\,\sigma^2_{\mathrm{a}})$ and the residuals were assumed to be distributed as $N(0,\sigma^2_{\mathrm{e}})$. The genetic variance σ^2_{a} and residual variance σ^2_{e} were estimated using ASReml (Gilmour et al. (2002)).

To avoid singularities in **G**, it was weighted by **A** as follows: $\mathbf{G}\omega = \omega\mathbf{G} + (1 - \omega)\mathbf{A}$ (VanRaden (2008)), with a weighting factor (ω) of 0.99.

Results and Discussion

The average LD between adjacent markers measured as r^2 (Hill and Robertson (1968)) was 0.41. Minor allele frequency was 0.29. Gene content for 1000 animals was predicted with an average accuracy of 0.58.

The accuracies of GEBV and regression coefficients are presented in Table 1. In general, adding animals with predicted genotypes to the reference population did not increase the accuracy of GEBV. However, for the lowest heritability an insignificant increase for the juveniles was observed. The low accuracy of GEBV for animals with predicted genotypes and for juveniles, observed in this study, is likely a consequence of the limited accuracy of predicted genotypes. The accuracy of predicting the genotypes could be improved by choosing animals with more genotyped relatives, especially genotyped offspring. Alternatively, instead of predicting genotypes, the **A** matrix could be enriched with the genomic information as proposed by Legarra et al. (2009) and used to estimate GEBV.

Comparison of the estimates for BLUP (scenario 4) and G-BLUP in scenario 2 showed that BLUP yielded accuracies lower than G-BLUP for the juvenile animals, as also was reported by Meuwissen et al. (2001). The difference in accuracy for juvenile animals comparing G-BLUP with traditional BLUP was increasing with heritability in contrast to Mulder et al. (2010) for gene-assisted breeding value estimation.

Table 1: Accuracies (Acc.) and regression coefficients (Reg.) of genomic estimated breeding value for groups of 1000 first, additional and juvenile animals for heritability of 0.3 (0.89 DYD), 0.05 (0.56 DYD) and 0.01 (0.2 DYD) for all scenarios

h^2	Scenario[1]	First 1000 anim.		Additional		Juvenile anim.	
		Acc.	Reg.	Acc.	Reg.	Acc.	Reg.
0.3	1	0.96^{A}[2]	1.02	-	-	0.84^{H}	1.01
	2	0.93^{B}	1.15	0.65^{E}	1.27	0.80^{I}	1.19
	3	0.96^{C}	1.02	0.96^{F}	1.02	0.90^{J}	1.01
	4	0.95^{D}	1.01	0.95^{G}	1.00	0.57^{K}	1.04
0.05	1	0.85^{A}	1.02	-	-	0.70^{G}	0.99
	2	0.85^{A}	1.07	0.62^{D}	1.23	0.70^{G}	1.01
	3	0.88^{B}	1.00	0.87^{E}	1.00	0.79^{H}	0.97
	4	0.81^{C}	0.99	0.79^{F}	1.00	0.48^{I}	1.00
0.01	1	0.65^{A}	1.03	-	-	0.52^{H}	1.01
	2	0.69^{B}	1.02	0.51^{E}	1.18	0.54^{H}	0.98
	3	0.72^{C}	0.99	0.70^{F}	0.99	0.60^{I}	0.94
	4	0.50^{D}	1.00	0.57^{G}	1.00	0.33^{J}	0.91

[1] scenarios: 1 – Scenario consisting of 1000 genotyped animals; 2 – Scenario consisting of 1000 genotyped and 1000 ungenotyped animals; 3 – Scenario consisting of 2000 genotyped animals; 4 – Scenario consisting of 2000 genotyped animals analyzed with use of traditional BLUP.

[2] Values with identical superscript did not differ significantly (P > 0.05); the significance of the differences among scenarios was examined within one heritability; standard errors of 10 replicates ranged from 0 to 0.012.

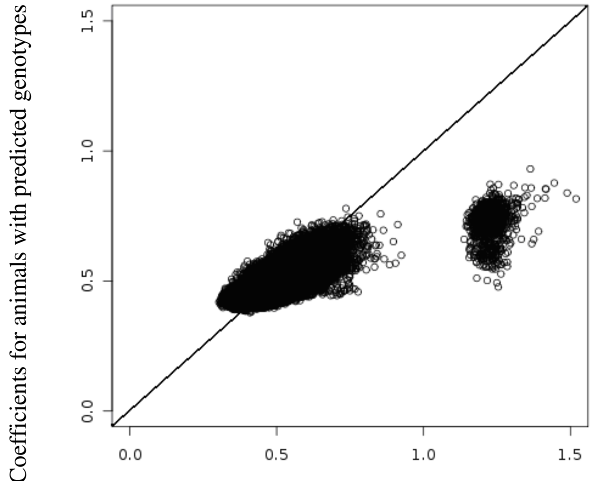

Figure 1: Coefficients for the animals with predicted genotypes versus coefficients for the genotyped animals

The low accuracy of GEBV in scenario 2 is most likely due to the limited accuracy of genotype prediction. Therefore, we compared the coefficients of the **G** matrix for animals with known or predicted genotype (Figure 1). The coefficients in the **G** matrix showed some inaccuracy, but also appeared to be on a different scale than the expected coefficients, especially the diagonal elements. This apparent bias of the **G** coefficients may be relieved by adjusting the elements of the **G** matrix, e.g. by regression of **G** on **A**-coefficients (VanRaden (2008)).

Conclusion

This study showed that inclusion of ungenotyped animals to the reference population tended to increase GEBV accuracies for juvenile animals when the heritability is low. Insignificance of the increase was most likely due to the low accuracy of predicted genotypes and, as a consequence, inaccuracy of the **G** matrix.

Acknowledgments

MP acknowledges the financial support of the Koepon Foundation (Amersfoort, The Netherlands). HM 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

References

Gengler, N., Mayeres, P., and Szydlowski, M. (2007). *Animal.* 1:21-28.

Gilmour, A. R., Gogel, B. J., Cullis, B. R. et al. (2002). VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Hill, W. and Robertson, A. (1968). Theoretical and Applied Genetics. 38:226-231.

Legarra, A., Aguilar, I., and Misztal, I. (2009). J. Dairy Sci. 92:4656-4663.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Genetics. 157:1819-1829.

Mrode, R. (2005). CABI, Wallingford, UK.

Mulder, H. A., Meuwissen, T. H. E., Calus, M. P. L. et al. (2010). Animal. 4 9-19.

Schaeffer, L. R. (2006). Journal of Animal Breeding and Genetics. 123:218-223.

VanRaden, P. M. (2008). J. Dairy Sci. 91:4414-4423.