J. Dairy Sci. 100:8277–8281
https://doi.org/10.3168/jds.2017-12734

Short communication: The role of genotypes from animals without phenotypes in single-step genomic evaluations

T. Shabalina,[1] E. C. G. Pimentel,[1,2] C. Edel, L. Plieschke, R. Emmerling, and K.-U. Götz
Institute of Animal Breeding, Bavarian State Research Center for Agriculture, 85586 Grub, Germany

ABSTRACT

In a 2-step genomic system, genotypes of animals without phenotypes do not influence genomic prediction of other animals, but that might not be the case in single-step systems. We investigated the effects of including genotypes from culled bulls on the reliability of genomic predictions from single-step evaluations. Four scenarios with a constant amount of phenotypic information and increasing numbers of genotypes from culled bulls were simulated and compared with respect to prediction reliability. With increasing numbers of genotyped culled bulls, there was a corresponding increase in prediction reliability. For instance, in our simulation scenario the reliability for selection candidates was twice as large when all culled bulls from the last 4 generations were included in the analysis. Single-step evaluations imply the imputation of all nongenotyped animals in the pedigree. We showed that this imputation was increasingly more accurate as increasingly more genotypic information from the culled bulls was taken into account. This resulted in higher prediction reliabilities. The extent of the benefit from including genotypes from culled bulls might be more relevant for small populations with low levels of reliabilities.
Key words: genomic breeding value, imputation, numerator relationship, SNP

Short Communication

The genomic selection method proposed by Meuwissen et al. (2001) consisted of a calibration step using information from genotyped animals with phenotypes and a prediction step for genotyped animals without phenotypes. This 2-step model can be formulated as a function of marker effects or of animal effects, with covariance structure described by the genomic relationship matrix **G** (VanRaden, 2008). The latter has been usually referred to as genomic BLUP. Legarra et al. (2009) and Christensen and Lund (2010) proposed a single-step model [single-step genomic BLUP (**ssGBLUP**)] for genomic evaluation in which all available information from genotypes, phenotypes, and pedigree is simultaneously used. This single-step model is expected to provide unbiased predictions even with selection and nonrandom mating (Fernando et al., 2014).

Due to decreasing genotyping costs, the number of animals being genotyped has been constantly increasing, which can lead to some computational challenges. This may be an issue, in particular, in single-step evaluations in which the system sizes are at least as large as the number of animals in the pedigree. Attempts to overcome these challenges have been proposed. Misztal et al. (2014) introduced a method for approximating the inverse of **G**, which is needed in ssGBLUP. Lourenco et al. (2014) examined the effect of reducing the number of generations used in ssGBLUP and reported that truncating old data may reduce computational requirements without decreasing reliability of prediction.

In a 2-step model, genotyped animals without phenotypes do not take part in the calibration step and, therefore, have no influence on predicted breeding values of other animals. In practice, this would mean that once young genotyped bulls have been culled their genotypes can be neglected in further runs of a 2-step genomic evaluation system. If such culled genotyped animals can also be neglected in ssGBLUP without affecting the prediction reliability for other animals, this would help reducing computational costs. However, Legarra et al. (2014) suggested that such animals should not be eliminated in ssGBLUP because they may change pedigree relationships across other animals. The objective of the present study was to investigate whether, to which extent, and how genotyped animals without phenotypes can influence the reliability of prediction in ssGBLUP.

To investigate these questions we simulated data using the software QMSim (Sargolzaei and Schenkel, 2009). Simulation was done in a way to mimic the structure and linkage disequilibrium found in the German-Austrian Fleckvieh population, as described in Plieschke et al. (2016). Briefly, a historical population of 2,000 unrelated animals with equal sex ratio was

Received February 15, 2017.
Accepted June 15, 2017.
[1]These authors contributed equally to this manuscript.
[2]Corresponding author: Eduardo.Pimentel@LfL.bayern.de

Table 1. Description of the amount of genotypic information in the 4 scenarios

	Number of genotyped males			
Generation	Scenario 1	Scenario 2	Scenario 3	Scenario 4[1]
5	1,050	1,050	15,000	1,050 + 1,050
6	1,050	1,050	15,000	1,050 + 1,050
7	1,050	1,050	15,000	1,050 + 1,050
8	1,050	15,000	15,000	1,050 + 1,050
9	3,000	3,000	3,000	3,000
Total	7,200	21,150	63,000	11,400

[1]The 1,050 highest and 1,050 lowest ranking bulls from generations 5 to 8.

generated and randomly mated for 2,500 generations. To create a linkage disequilibrium structure similar to the one observed in Fleckvieh, a bottleneck was introduced in generation 2,501 by reducing the number of breeding animals to 150, which approximately corresponds to the effective population size estimated for Fleckvieh (Pausch et al., 2013). Afterward the number of breeding animals was increased to 30,000 dams and 1,500 sires. These were the founder animals (generation 0) of the recent population, which was propagated for another 9 generations. In each of them, 15,000 female and 15,000 male offspring were produced. Generations were overlapping and in every generation 30% of the dams (n = 9,000) and 70% of the sires (n = 1,050) were replaced. These replacement ratios are quite similar to the situation observed in the real Fleckvieh population. Breeding animals were selected based on their EBV, which were calculated by QMSim. The reliabilities of these EBV were approximately 0.6, which resulted from the availability of information on parent average plus a single phenotypic record of the own animal. The simulated genome consisted of 30 chromosomes with 100 centimorgans. Single nucleotide polymorphisms were evenly distributed and QTL were randomly distributed across the genome (49,800 SNP and 900 QTL in total). To ensure sufficient quality of the genotypic data, SNP that significantly deviated from Hardy–Weinberg equilibrium ($P < 10^{-5}$) and SNP with minor allele frequency lower than 0.02 were excluded. After quality control, about 37,500 SNP and 700 QTL remained in the data set. True breeding values were simulated as the sum of QTL effects, and phenotypes were generated considering a heritability of 0.4. The polygenic nature of the trait was ensured by the relatively high number of QTL whose effects were drawn from a uniform distribution; we did this to prevent the occurrence of a few isolated large QTL effects. Despite the fact that QMSim simulated phenotypes for all animals, we only used phenotypes of females to depict a situation observed in dairy cattle. Simulation was replicated 5 times and we also ran one repetition in which replacement animals were selected at random.

To show the influence of genotypes without phenotypes, we assigned the simulated data to 4 different scenarios containing an increasing number of culled genotyped males. Phenotypic information did not change across scenarios and consisted of phenotypes of the cows from generations 3 to 8 (15,000 cows with phenotypes per generation, summing to a total of 90,000 phenotypic records). The numbers of genotyped male animals from each generation for the 4 different scenarios are shown in Table 1.

The first scenario (**S1**) contained genotypes of all the bulls with daughter performance from generations 5 to 7, of the top young bulls considered being selected but still without daughter performance from generation 8, and of the young selection candidates from generation 9. The amount of genotypic information in this reference scenario depicts a typical situation in dairy genomic programs running 2-step evaluations, in which only genotypes of calibration bulls, genomic young bulls, and candidates are used. The second scenario (**S2**) additionally contained genotypes of all culled young bulls from generation 8. In the third scenario (**S3**), all culled bulls from generations 5 to 7 were also included. The fourth scenario (**S4**) was an alternative to S3 in which not all, but only the subset of the highest and lowest ranking bulls from generations 5 to 8 were included. This was done to check if the addition of the lowest-ranking nonselected animals would make the additional genotype pool representative enough so that at least some computational demands could be saved.

We estimated genomic EBV for all scenarios with ssGBLUP using the software package MiX99 Release VIII/2015 (Lidauer et al., 2015) for solving the mixed model equations. The statistical model included a random animal effect and an overall mean as fixed effect. The system of equations to solve is similar to Henderson's mixed model equations for an animal model, except that the covariance structure of the animal effect is described by a matrix $\mathbf{H}$ instead of the usual numerator relationship matrix $\mathbf{A}$. According to Aguilar et al. (2010) and Christensen and Lund (2010), the inverse of $\mathbf{H}$ has the following form: