Size of required reference population updates to achieve constant genomic prediction accuracy across generations

***M. Pszczola*[*], T. Strabel[*], and M.P.L. Calus[†]**
[*]Poznan University of Life Sciences, Poland, [†]Wageningen UR Livestock Research, The Netherlands

ABSTRACT: High accuracy of estimated breeding values is crucial for achieving high genetic progress. The accuracy of genomic breeding values drops if the reference population is used over generations without supplementing it with new animals. The goal of this study was to investigate how many animals per generation need to be added to the reference population to keep the accuracy at a constant level across generations. On average the accuracy dropped by 0.07 when moving from first to second generation. After adding 25% of the initial reference population size the accuracy returned to its initial value. The required number of animals that were added to the reference population varied substantially illustrating that probably there are better strategies than adding animals at random. A possible solution is to consider the relationships between the animals used to update reference population and the selection candidates.
Keywords: dairy cattle; reference population; accuracy; relationships

Introduction

High accuracy of estimated breeding values is crucial for achieving high genetic progress. Establishing the reference population is an important step in a breeding program in which genomic evaluation is implemented. An optimally designed reference population enables maximizing the accuracy for the given population (Pszczola et al., 2012). Once established, the reference population can be used to evaluate the selection candidates. However, if the reference population is used over generations without supplementing it with new animals, the breeding value accuracy drops (Calus, 2010; Wolc et al., 2011). As the breeding value accuracy depends on the relationships between the reference population and the selection candidates (Habier et al., 2010; Clark et al., 2012; Pszczola et al., 2012), one reason of this drop in accuracy is the decay in the relationships. Consequently, it is important to update the reference population with animals from the next generations to maintain the accuracy constant at its initial level. An important question is what is the minimum number of animals that needs to be added to the reference population. The goal of this study was to investigate how many animals per generation need to be added to the reference population every generation to keep the accuracy at a constant level across generations. This minimum required number of animals to update the reference population is specifically investigated for a scenario with a novel trait for which only a small reference population with own phenotypes. Examples of such a trait are methane emission or dry matter intake in dairy cattle.

Materials and Methods

Data. The dataset used in this study was simulated using QMSim software (Sargolzaei and Schenkel, 2009) to mimic a historic dairy cattle population and a modern population under selection. The values of effective population size (*Ne*) reflected different points in the history of the US and Canadian Holstein cattle (Schenkel et al., 2009). The initial generation of the modern population consisted of 5,025 animals of which 5,000 were females and 25 were males. In the initial generation of the modern population 25 males were mated to 5,000 females. This resulted in 2,500 males and 2,500 females as base for the next generations. Progeny sex ratio was 0.5 and the replacement ratio for females was 0.5, resulting in overlapping generations. Every generation, the 25 best males were mated to 5,000 females. The first 10 generations were used to establish Bulmer-equilibrium. The simulation was replicated 10 times.

Genome. The simulated genome consisted of 29 autosomes with a total genome length of 2,333 cM. The initial 46,660 markers and 7,250 QTL were evenly spaced across the genome. The simulation of the historic populations resulted in 43,256 segregating markers and 6,734 QTL.

Phenotypes. To simulate a novel trait that is not included in the breeding goal, but is affected indirectly by selection for correlated traits, we simulated two genetically correlated traits. The first trait resembled a breeding goal trait under the selection pressure and was simulated using the QMSim software. The assumed heritability was 0.25. The second trait mimicking the novel trait was simulated using the output from QMSim. The second trait was assumed to be genetically correlated with the trait under selection and to have a heritability of 0.15. The assumed genetic correlation was 0.25 and the two traits shared only half of the QTL affecting them. True breeding values for both traits were obtained by summing the QTL effects and phenotypes were simulated by adding a random residual term to the true breeding value. Only the second trait was subject for further analyses.

Reference population and selection candidates. The reference population and selection candidates were sampled randomly from the 2,500 females available per generation. The initial reference population consisted of 2,000 animals and the remaining 500 animals were considered to be the selection candidates. This set of the animals was used to establish the initial accuracy.

Analyses. The initial reference population of 2,000 cows was used to estimate variance components. The following animal model was fitted using ASReml 3.0 (Gilmour et al., 2009):

$$y_i = \mu_j + animal_j + e_i, \quad [1]$$

where y_j is the phenotypic record of animal j, μ_j is the overall mean, $animal_j$ is the random genomic effect of animal j and e_i is a random residual term. In the analyses we used the genomic relationship matrix (**G**) created with the first formula described by VanRaden (2008) using current allele frequencies. The estimated variance components were then used to estimate genomic breeding values in BLUP analyses performed using model 1 fitted in ASReml 3.0 (Gilmour et al., 2009).

Reference population update. In each of the following 4 generations the reference population from generation *n* was updated by 100 randomly sampled animals by *m* times, until the accuracy for the selection candidates from generation *n+1* reached the initial accuracy (see Figure 1). Once established, the initial accuracy was kept constant across generations. When the initial accuracy was reached the increased reference population was used to evaluate the selection candidates from the next generation. In the last round, the reference population consisted of animals from generations 1 to 4 and the last evaluated selection candidates were from generation 5.

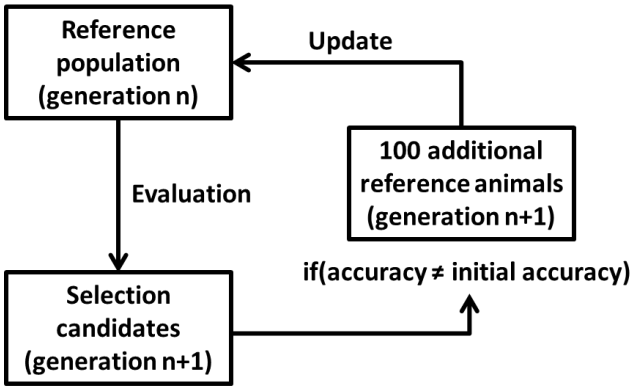

Figure 1: Scheme of updating the reference population with new animals .

Results and Discussion

The average initial accuracy was on average 0.61 and ranged from 0.57 to 0.66 across the replicates. The average accuracy when moving from first to the next generation dropped to 0.54, ranging from 0.48 to 0.63 across replicates. The average drop in the accuracy in the first generation (0.07) was generally higher than the drop in the first generation reported by Wolc et al. (2011), but similar to Pszczola et al. (2012). The mean increase per update of the reference population equaled 0.02 and was consistent over generations (see Table 1), however, differences across replicates were quite high and ranged from 0.003 to 0.033. Big variation among scenarios was also reflected in the number of animals required to be added to the reference population to re-gain the initial accuracy. On average about 475 animals were required (see Table 2). However, between replicates, in an extreme cases the update was not necessary or the assumed limit of the update of 1,000 individuals was not enough to re-gain the accuracy. The differences among scenarios might be due to sampling of the animals used to update the reference population. Therefore probably an optimized method to update the reference population with new animals is desired than choosing the animals at random. For example, Pszczola et al. (2012) showed that the optimally designed reference population consists of animals that are minimally related to each other and maximally related to the selection candidates. Also, other studies showed that closer relationship between the reference population and the selection candidates leads to higher accuracy (Habier et al., 2010; Wolc et al., 2011; Clark et al., 2012; Wientjes et al., 2013). Since close relationships between the reference population and selection candidates are important, animals used to update the reference population should be closely related to the potential selection candidates. Including such animals first, is expected to lead to a faster increase of the accuracy due to the update, and the initial accuracy could be reached earlier. Possibly, taking into account the relationships between the animals added to the reference population and the selection candidates could also lead to more consistent results.

Table 1: The average increase of the accuracy per 100 animals added to the reference population across the analyzed generations together with S.E. across replicates.

	Generation			
	1	2	3	4
Mean	0.017	0.014	0.017	0.014
S.E.	0.0037	0.0038	0.005	0.0045

Table 2: The number of animals required to re-gain the initial accuracy over the analyzed generations averaged over the 10 replicates.

	Generation			
	1	2	3	4
Mean	650	400	510	340
S.E.	116	125	125	111

Conclusion

To maintain the initial level of the accuracy, the reference population needs to be updated. Not only the number of the animals added to the reference population is important but also a selection of which animals should be added, as there were large differences in impact on the accuracy across different updates. Further investigation in the methods that will take into account the relationships between the animals being added to the reference

population and the selection candidates is needed. Such method could lead to decrease in the number of the animals added per generation and is expected to give more consistent results.

Literature Cited

Calus, M.P.L. (2010). Animal 4:157-164.

Clark, S., Hickey J., Daetwyler H. et al. (2012). Genet. Sel. Evol. 44:4.

Gilmour, A.R., Gogel B.J., Cullis B.R. et al. (2009). VSN International Ltd, Hemel Hempstead, UK.

Habier, D., Tetens J., Seefried F.-R. et al. (2010). Genet. Sel. Evol. 42:5-17.

Pszczola, M., Strabel T., Mulder H.A. et al.(2012). J. Dairy Sci. 95:389-400.

Sargolzaei, M. and Schenkel F.S. (2009). Bioinformatics 25:680-681.

Schenkel, F., Sargolzaei M., Kistemaker G. et al. (2009). Interbull Bulletin 39:51-58.

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

Wientjes, Y.C.J., Veerkamp R.F., Calus M.P.L. (2013). Genetics 193:621-631.

Wolc, A., Arango J., Settar P. (2011). Genet. Sel. Evol. 43:23.