J. Dairy Sci. 100:1–7
https://doi.org/10.3168/jds.2016-11335

Accuracy of breeding values in small genotyped populations using different sources of external information—A simulation study

S. Andonov,*[1] D. A. L. Lourenco,† B. O. Fragomeni,† Y. Masuda,† I. Pocrnic,† S. Tsuruta,† and I. Misztal†
*Faculty of Agricultural Sciences and Food, University of Ss. Cyril and Methodius, PO Box 297, 1000 Skopje, Macedonia
†Department of Animal and Dairy Science, University of Georgia, Athens 30602

ABSTRACT

Genetically linked small and large dairy cattle populations were simulated to test the effect of different sources of information from foreign populations on the accuracy of predicting breeding values for young animals in a small population. A large dairy cattle population (P_L) with >20 generations was simulated, and a small subpopulation (P_S) with 3 generations was formed as a related population, including phenotypes and genomic information. Predicted breeding values for young animals in the small population were calculated using BLUP and single-step genomic BLUP (ssGBLUP) in 4 different scenarios: (S1) 3,166 phenotypes, 22,855 pedigree animals, and 1,000 to 6,000 genotypes for P_S; (S2) S1 plus genomic estimated breeding value (GEBV) for 4,475 sires from P_L as external information; (S3) S1 plus 221,580 phenotypes, 402,829 pedigree animals, and 53,558 genotypes for P_L; and (S4) single nucleotide polymorphism (SNP) effects calculated based on P_L data. The ability to predict true breeding value was assessed in the youngest third of the genotyped animals in the small population. When data only from the small population were used and 1,000 animals were genotyped, the accuracy of GEBV was only 1 point greater than the estimated breeding value accuracy (0.32 vs. 0.31). Adding external GEBV for sires from P_L did not considerably increase accuracy (0.33 vs. 0.32 in S1). Combining phenotypes, pedigree, and genotypes for P_S and P_L was beneficial for predicting accuracy of GEBV in the small population, and the prediction accuracy of GEBV in this scenario was 0.38 compared with 0.31 from estimated breeding values. When SNP effects from P_L were used to predict GEBV for young genotyped animals from P_S, accuracy was greatest (0.56). With 6,000 genotyped animal in P_S, accuracy was greatest (0.61) with the combined populations. In a small population with few genotypes, the highest accuracy of evaluation may be obtained by using SNP effects derived from a related large population.

Key words: dairy cattle, genomic breeding value, prediction accuracy, small population

INTRODUCTION

National dairy cattle breeding programs of 30 countries have become comparable because of the use of the same superior animals and participation in the multitrait, across-country evaluation (Schaeffer, 1994) of the Interbull Centre (Uppsala, Sweden). In countries with fairly small Holstein populations, population maintenance and progress rely on the continuous import of genetic material (pregnant heifers, semen, and embryos). Despite importation, such countries still have an interest in national breeding programs. Whereas reliabilities from evaluations with small populations are low, they can benefit from using evaluations from large reference populations, which have more traits recorded and higher reliabilities (Wiggans et al., 2011).

Genomic selection can provide increased reliability of predictions, increased genetic gain, and reduced generation interval (Schaeffer, 2006). Use of genomic selection for livestock has become widespread because of the availability of cost-effective genotyping. Adding genomic to pedigree and phenotypic information leads to increased accuracy of genomic EBV (**GEBV**), particularly when the training population has a large set of genomic and phenotypic data (Hayes et al., 2009; VanRaden et al., 2009; Lund et al., 2011). In addition, the genetic merit of young sires can be predicted with high accuracy without progeny performance. As a consequence, breeding organizations have begun to exchange genomic information for sires in addition to their EBV (Cromie et al., 2010). Members of the EuroGenomics consortium have already benefited from exchange of genomic data in GEBV estimation of sires used across countries and in genetic predictions for young animals without progeny records (Patry, 2015). In such evaluations, all available sources of information

---

Received April 19, 2016.
Accepted September 12, 2016.
[1] Corresponding author: sreten_andonov@yahoo.com or sandonov@zf.ukim.edu.mk

(e.g., relationship to common ancestors, phenotypes, and genotypes) are included and require robust and reliable methodology.

Information from large populations can be incorporated in several ways. First, EBV or GEBV from external data can be used as priors for BLUP, either in nongenomic (Legarra et al., 2007) or genomic context (Přibyl et al., 2014; Vandenplas et al., 2014; VanRaden et al., 2014). Second, if complete data sharing agreements are available, data from multiple countries can be combined. Use of all information including female genotypes is simplified by the use of single-step genomic BLUP (**ssGBLUP**; Aguilar et al., 2010; Legarra et al., 2014). It benefits from the inclusion of all relatives of nongenotyped and genotyped animals and their performance without losing information; an additional benefit is the use of the same models as in traditional evaluations. Initially, ssGBLUP could not account for >150,000 genotyped animals, but that limitation has been overcome through additional algorithms developed by Misztal et al. (2014). The so-called algorithm for proven and young was recently used in the evaluation of Holstein data with more than 7 million phenotypes and almost 570,000 genotypes (Masuda et al., 2016). Third, GEBV for genotyped animals in a small population can be computed by using SNP effects derived from a large population (Council on Dairy Cattle Breeding, 2016). Whereas GEBV from small populations require blending of SNP-derived values with parent average, that blending is no longer needed when SNP effects are derived from a large number of genotypes (Lourenco et al., 2015). In the extreme, SNP effects can be backsolved from genotypes and GEBV.

The different socio-economic and environmental conditions in different countries with small Holstein populations result in constraints in designing a breeding program that is genetically effective and economically reasonable. Even though genotyping cost has been decreasing, it is still an unaffordable and questionable investment for countries with small populations with limited economic power. The objective of this study was to investigate, by simulation, the accuracy of evaluation on a small population with different approaches for incorporating information from a large related population.

MATERIALS AND METHODS

Simulation

To mimic a Holstein population, data were generated with QMSim simulation software (Sargolzaei and Schenkel, 2009); 2 linked populations were simulated in 5 replicates. To start, a large population ($\mathbf{P_L}$) of 400,000 animals over 20 generations was simulated, and then a small population ($\mathbf{P_S}$) of 6,000 animals over only 3 generations and partially related to P_L was simulated. To produce a realistic level of linkage disequilibrium, 800 generations of a historical population were generated with initially 3,000 animals, increasing to 40,000 at generation 500, and then decreasing to 30,000 by the last generation; one-tenth of animals were sires. The P_L was founded by 2,000 sires and 20,000 dams from the historical population; single records were generated for dams, and heritability was set to 0.3. The replacement rate for sires and dams was 90 and 30%, respectively; the selection criterion was based on high EBV. Only 1 calf was allowed for each calving. For P_S, the first generation was founded by 150 sires and 2,000 dams from generation 18 of P_L; for the second generation, an additional 100 sires were used from generation 19 of P_L to ensure strong genetic ties between P_L and P_S. The structure of P_S was similar to P_L except for less intensive culling of sires (50%) and random selection criterion.

Genotypes were simulated for 45,000 biallelic SNP markers distributed along 29 chromosomes with a total length of 2,319 cM, which mimicked the bovine genome without sex chromosomes. A total of 450 biallelic and randomly distributed QTL affected the trait, with effects sampled from a normal distribution. Genomic information was generated for generations 14 through 20 of P_L and all 3 generations of P_S. Because of the considerable amount of genomic information generated for P_L, only genotypes for sires and dams with more than 2 offspring were kept, thus reducing the computational effort in analysis that included P_L genotypes. Mean linkage disequilibrium ranged from 0.3 to 0.4 for both populations among replicates.

Scenarios

Four different scenarios were defined according to the source of information used for the evaluation (Table 1). The first scenario (**S1**) mimicked a local evaluation where, on average, 3,166 phenotypes, 22,885 pedigree animals, and 1,000 genotypes were available only for animals in P_S; in this scenario, we indicate the output GEBV as P_S GEBV. The second scenario (**S2**) used all the information for S1 plus external GEBV from 4,475 P_L sires of animals in P_S, which was used as prior information for the local evaluation (Legarra et al., 2007); in this scenario, we indicate the output GEBV as P_S+ GEBV. The next scenario (**S3**) mimicked an ideal situation where phenotypes, pedigree, and genotypes from a large population are available to be included in the