J. Dairy Sci. 97:3918–3929
http://dx.doi.org/10.3168/jds.2013-7761

Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population

C. Hozé,*†[1] S. Fritz,† F. Phocas,* D. Boichard,* V. Ducrocq,* and P. Croiseau*
*Institut National de la Recherche Agronomique (INRA), UMR 1313, Génétique Animale et Biologie Intégrative (GABI), 78350 Jouy-en-Josas, France
†Union Nationales des Coopératives d'Élevages et d'Insémination Animales (UNCEIA), 149 rue de Bercy, 75012 Paris, France

ABSTRACT

Single-breed genomic selection (GS) based on medium single nucleotide polymorphism (SNP) density (~50,000; 50K) is now routinely implemented in several large cattle breeds. However, building large enough reference populations remains a challenge for many medium or small breeds. The high-density BovineHD BeadChip (HD chip; Illumina Inc., San Diego, CA) containing 777,609 SNP developed in 2010 is characterized by short-distance linkage disequilibrium expected to be maintained across breeds. Therefore, combining reference populations can be envisioned. A population of 1,869 influential ancestors from 3 dairy breeds (Holstein, Montbéliarde, and Normande) was genotyped with the HD chip. Using this sample, 50K genotypes were imputed within breed to high-density genotypes, leading to a large HD reference population. This population was used to develop a multi-breed genomic evaluation. The goal of this paper was to investigate the gain of multi-breed genomic evaluation for a small breed. The advantage of using a large breed (Normande in the present study) to mimic a small breed is the large potential validation population to compare alternative genomic selection approaches more reliably. In the Normande breed, 3 training sets were defined with 1,597, 404, and 198 bulls, and a unique validation set included the 394 youngest bulls. For each training set, estimated breeding values (EBV) were computed using pedigree-based BLUP, single-breed BayesC, or multi-breed BayesC for which the reference population was formed by any of the Normande training data sets and 4,989 Holstein and 1,788 Montbéliarde bulls. Phenotypes were standardized by within-breed genetic standard deviation, the proportion of polygenic variance was set to 30%, and the estimated number of SNP with a nonzero effect was about 7,000. The 2 genomic selection (GS) approaches were performed using either the 50K or HD genotypes. The correlations between EBV and observed daughter yield deviations (DYD) were computed for 6 traits and using the different prediction approaches. Compared with pedigree-based BLUP, the average gain in accuracy with GS in small populations was 0.057 for the single-breed and 0.086 for multi-breed approach. This gain was up to 0.193 and 0.209, respectively, with the large reference population. Improvement of EBV prediction due to the multi-breed evaluation was higher for animals not closely related to the reference population. In the case of a breed with a small reference population size, the increase in correlation due to multi-breed GS was 0.141 for bulls without their sire in reference population compared with 0.016 for bulls with their sire in reference population. These results demonstrate that multi-breed GS can contribute to increase genomic evaluation accuracy in small breeds.

Key words: multi-breed genomic selection, dairy cattle, high-density chip

INTRODUCTION

Genomic selection has been implemented in many countries. To date, more than 16 countries apply genomic information for dairy cattle breeding (Nilforooshan et al., 2010; Eggen, 2012). However, the accuracy of genomic breeding values depends mainly on the size of the reference population (Hayes and Goddard, 2008; Goddard, 2009) and therefore, genomic evaluations are mainly implemented in large breeds. In small breeds, only a small number of progeny-tested bulls is available and assembling a large reference population consisting of animals with accurate phenotypes is challenging. In this context, a potential solution is to combine reference populations from different breeds and develop multi-breed genomic selection (**GS**).

Such an approach requires conserved linkage disequilibrium (**LD**) across breeds to maintain the association between SNP and QTL. Studies on real data show that the association between marker alleles is maintained for SNP <10 kb apart (Gautier et al., 2007; de Roos et al., 2008), a condition not fulfilled with the classically used BovineSNP50 BeadChip (**50K**, ~50,000 SNP; Illumina

Received November 25, 2013.
Accepted February 25, 2014.
[1]Corresponding author: chris.hoze@jouy.inra.fr

Inc., San Diego, CA). A simulation study by de Roos et al. (2009) confirmed these results and found that the most accurate genomic prediction was achieved when all reference populations were combined and a chip containing more than 300,000 SNP was used. Therefore, the BovineHD BeadChip (**HD**; Illumina Inc.), developed in 2010 and containing ~777,000 SNP, should be sufficiently dense to allow for efficient multi-breed GS.

Given the large number of animals already genotyped on the 50K chip, regenotyping reference populations on this HD chip is not economically justified. Imputation studies reported an observed allelic imputation error rate <1% when 50K genotypes were imputed to HD (Su et al., 2012; Hozé et al., 2013; Pausch et al., 2013; Schrooten et al., 2014). These low error rates are expected to have a minor effect on the reliability of GS (Dassonneville et al., 2011; Mulder et al., 2012). Imputation of HD reference populations from 50K genotypes and investigations on the benefit of the genomic evaluation are therefore possible.

Until now, studies on the 50K and HD panels showed limited gain in accuracy when comparing multi-breed to single-breed GS (Hayes et al., 2009; Erbe et al., 2012). However, when comparing methods, these researchers confirmed the advantage of Bayesian approaches compared with genomic BLUP for EBV estimation (Hayes et al., 2009; Erbe et al., 2012); they stated that setting a large proportion of SNP effects to zero is necessary to take advantage of the density of the HD chip (Erbe et al., 2012). This conclusion is in agreement with conserved QTL-marker association at small distances only. Furthermore, adding a polygenic component avoids spurious SNP-QTL associations due to pedigree relationship (Solberg et al., 2009; Liu et al., 2011) and helps to select QTL with rare alleles, small effects, or both (Calus and Veerkamp, 2007; Goddard, 2009). Inclusion of a polygenic component also increases the accuracy of genomic EBV (**GEBV**) prediction and allows for regression slopes closer to 1 [M. Gunia (Institut National de la Recherche Agronomique Génétique Animale et Biologie Intégrative (INRA GABI), Jouy-en-Josas, France), R. Saintilan (INRA GABI; UNCEIA, Paris, France), E. Venot (INRA GABI), C. Hozé, M. N. Fouilloux (Institut de l'Élevage, Paris, France), and F. Phocas; unpublished data].

Within-breed, genomic evaluation relies on short distance QTL-SNP associations and on long-distance LD due to relationships. We assumed here that in a multi-breed situation, across-breed information is brought only by QTL-SNP associations shared across breeds; therefore, focusing on them should avoid detecting SNP associated with genetic background of the breed. BayesC (Kizilkaya et al., 2010) and BayesCπ (Habier et al., 2011) approaches have been widely used in GS programs and allow setting a proportion (π) of SNP with a zero effect. Therefore, these approaches were chosen here to compare accuracy of single-breed and multi-breed GS for a small reference population.

In small breeds, reference population size is limiting. Not only is the training set small but assessing the achieved accuracy is difficult because of the small validation population. If the validation population is enlarged, it would be at the expense of the training set and, therefore, at the expense of prediction accuracy (Erbe et al., 2010). Therefore, we chose here to use a large dairy breed to mimic a small breed and develop a multi-breed GS method. This strategy offers the opportunity to study several training population sizes to mimic either small or large breeds while using a unique reasonably large validation set. We first investigated the phenotypes used, the proportions of SNP with a nonzero effect, and the proportion of residual polygenic variance on a 50K basis. Then, we compared the predictive ability of genomic evaluation based on single-breed and multi-breed reference populations using the HD data set. Then, we assessed the benefit of using the HD chip for multi-breed analysis and investigated the effect of population structure on GS accuracy.

MATERIALS AND METHODS

Reference Population

Genotypes used in this study came from the French genomic evaluation. The biological tissues, either cryopreserved semen or blood samples, were provided by various commercial AI companies and breeder organizations in the framework of their breeding program activities. Therefore, no ethical approval was required for this study.

In total, 535, 527, and 773 influential bulls from the Normande (**NO**), Montbéliarde (**MO**), and Holstein (**HO**) breeds were genotyped with the Illumina Bovine HD BeadChip and were used to impute HD genotypes for animals of the French reference population genotyped with the 50K chip. Quality control was performed within breed on the HD and 50K genotypes, using the same criteria for both chips. Genotyped animals with a call rate <0.95 were removed from the analysis. Only markers mapped on the UMD3.1 assembly (http://bovinegenome.org/cgi-bin/gbrowse/bovine_UMD31/) covering the 29 bovine autosomes were used. Any SNP showing departure from Hardy-Weinberg equilibrium ($P < 0.0001$) or with more than 10% missing genotypes were removed. In addition, genotype consistency was checked using 1,838 animals that were genotyped on both chips, and 352 markers that were discordant for more than 1% of these animals were excluded.