J. Dairy Sci. 98:3478–3484
http://dx.doi.org/10.3168/jds.2014-9132

Effect of reference population size and available ancestor genotypes on imputation of Mexican Holstein genotypes[1]

A. García-Ruiz,*† F. J. Ruiz-Lopez,*†[2] G. R. Wiggans,‡ C. P. Van Tassell,‡ and H. H. Montaldo§
*Facultad de Estudios Superiores Cuautitlán, Facultad de Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, México, DF 04510, México
†Centro Nacional de Investigación en Fisiología y Mejoramiento Animal, Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Ajuchitlán, Querétaro 76280, México
‡Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350
§Facultad de Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, México, DF 04510, México

ABSTRACT

The effects of reference population size and the availability of information from genotyped ancestors on the accuracy of imputation of single nucleotide polymorphisms (SNP) were investigated for Mexican Holstein cattle. Three scenarios for reference population size were examined: (1) a local population of 2,011 genotyped Mexican Holsteins, (2) animals in scenario 1 plus 866 Holsteins in the US genotype database (GDB) with genotyped Mexican daughters, and (3) animals in scenario 1 and all US GDB Holsteins (338,073). Genotypes from 4 chip densities (2 low density, 1 mid density, and 1 high density) were imputed using findhap (version 3) to the 45,195 markers on the mid-density chip. Imputation success was determined by comparing the numbers of SNP with 1 or 2 alleles missing and the numbers of differently predicted SNP (conflicts) among the 3 scenarios. Imputation accuracy improved as chip density and numbers of genotyped ancestors increased, and the percentage of SNP with 1 missing allele was greater than that for 2 missing alleles for all scenarios. The largest numbers of conflicts were found between scenarios 1 and 3. The inclusion of information from direct ancestors (dam or sire) with US GDB genotypes in the imputation of Mexican Holstein genotypes increased imputation accuracy by 1 percentage point for low-density genotypes and by 0.5 percentage points for high-density genotypes, which was about half the gain found with information from all US GDB Holsteins. A larger reference population and the availability of genotyped ancestors improved imputation; animals with genotyped parents in a large reference population had higher imputation accuracy than those with no or few genotyped relatives in a small reference population. For small local populations, including genotypes from other related populations can aid in improving imputation accuracy.

Key words: imputation, Mexican Holstein, ancestor genotype, reference population

INTRODUCTION

Genomic selection is a recent technology that has enhanced genetic improvement in dairy cattle but its success relies on numerous factors. The most important are methodology for calculating marker effects (Hayes et al., 2009; Daetwyler et al., 2010; VanRaden et al., 2010), marker panel density (Solberg et al., 2008; Hayes et al., 2009), trait heritability (Hayes et al., 2009), and number of genotyped animals in the reference population (Hayes et al., 2009). To improve the reliability of genomic predictions, increasing the number of genotyped animals is more important than using higher density panels (VanRaden et al., 2010). In developing countries such as Mexico, genotyping is still expensive and high-density marker panels are generally more expensive than low-density panels. Using low-density panels is an alternative that may lead to more genotyped animals.

Because genomic technology has evolved rapidly, the number and sets of markers to be used for genomic prediction have changed over time (Hayes et al., 2009). To combine information based on different markers or marker densities, missing markers for animals genotyped with lower density panels are imputed from genotype information of relatives or from haplotypes of animals genotyped with higher density panels (Druet et al., 2010; VanRaden et al., 2010). Druet et al. (2010) showed that imputation efficiency is higher when the size of the reference population, marker density, and expected proportion of the genome inherited from the reference population are increased. Bouwman et al. (2014)

Received November 20, 2014.
Accepted February 2, 2015.
[1]The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the US Department of Agriculture or the Agricultural Research Service of any product or service to the exclusion of others that may be suitable.
[2]Corresponding author: ruiz.felipe@inifap.gob.mx

reported that the number of genotyped relatives has a direct effect on imputation accuracy for animals without genotypes, and Khatkar et al. (2012) demonstrated the importance of considering genotyped relatives when imputing low-density genotypes, particularly when the sire's genotype was not included in the reference genotypes. Although different imputation methods with high accuracy have been implemented (Browning and Browning, 2011; VanRaden et al., 2011; Hickey et al., 2012), the choice of the optimal method depends on population structure (Johnston et al., 2011).

Using imputation as part of genomic selection reduces genotyping costs and increases both the size of the reference population and the number of markers for which effects are estimated, which increases the reliability of genetic predictions and consequently the expected genetic improvement. Reducing genotyping cost also makes the technology more accessible to breeders (Berry and Kearney, 2011; VanRaden et al., 2011). For dairy cattle, reliability of genomic predictions using imputation varies according to trait and according to population size and structure. Reliability improvements of approximately 2 percentage points have been reported in simulation studies that included a set of 500,000 SNP per imputed genotype compared with a 50,000-marker subset (VanRaden et al., 2011).

Canadian and US dairy cattle genetics are widely used around the world. Using genotypes from those animals to impute local genotypes would increase the size of the local reference population considerably and could increase the accuracy of imputing genotypes, thus improving genomic evaluations in many countries. Such an improvement should be possible for Mexico, because Mexican Holstein breeders have depended heavily on US and Canadian genetics. The primary objective of this study was to determine the effect on imputation of Mexican Holstein genotypes from increasing the size of the Mexican Holstein reference population by adding Holsteins from the US genotype database. The effect of relatedness of animals in the reference population on imputation also was evaluated.

MATERIALS AND METHODS

Scenarios

Three imputation scenarios were defined based on source and number of genotyped animals. For scenario 1, only genotypes of a local population of 2,011 genotyped Mexican Holsteins were included. For scenario 2, genotypes of animals in scenario 1 plus genotypes of 866 Holsteins in the US genotype database (**GDB**) with genotyped Mexican daughters were included. For scenario 3, genotypes of animals in scenario 1 and all US GDB Holsteins available at the time of the study (338,073) were included.

Data

Genotypes. The genotyped Mexican Holstein population (scenario 1) included 1,971 cows and 40 sires. For cows, 183 were genotyped with the Illumina BovineLD BeadChip v1.1 (**6K**; Illumina Inc., 2013), 277 with the GeneSeek Genomic Profiler BeadChip v1 (**9K**; Neogen Corp., 2013a), 686 with the Illumina BovineSNP50 BeadChip v2 (**50K**; Illumina Inc., 2011), and 825 with the GeneSeek Genomic Profiler HD BeadChip (**77K**; Neogen Corp., 2013b). All Mexican sires had 50K genotypes. Of the US GDB Holsteins, 839 bulls and 47 cows had genotyped daughters in Mexico (scenario 2). All US GDB dams of genotyped Mexican daughters had 50K genotypes; the US GDB sires included 533 US, 270 Canadian, and 22 European bulls with 50K genotypes and 10 US and 4 Canadian bulls with 77K genotypes. For scenario 3, 338,073 US GDB genotypes were included.

Pedigrees. Two different pedigree files were used in the analysis: 27,625 animals for scenarios 1 and 2; 938,662 animals for scenario 3.

Imputation

Missing genotypes were predicted by combining population and pedigree haplotypes with findhap (version 3) software (VanRaden, 2015). The imputation goal for all scenarios was to fill in any missing genotypes for the 45,195 SNP from the 50K chip that were then being used in US genomic evaluations. Of those 45,195 SNP, 6,842 were included in 6K genotypes, 8,196 in 9K genotypes, and 28,048 in 77K genotypes, and a different number of SNP was imputed for each chip density: 38,353 for 6K genotypes, 36,999 for 9K genotypes, and 17,147 for 77K genotypes. For imputation, genotypes were first coded as 0 = BB, 1 = AB, 2 = AA, or 5 = both alleles unknown. Then, SNP genotypes for each scenario were imputed, and the results were compared using SAS software (version 9.3; SAS Institute Inc., Cary NC).

A SNP genotype cannot always be determined through imputation. If only 1 parental allele could be determined (i.e., B and unknown allele or A and unknown allele), then the SNP genotype was designated as M1. If neither parental contribution could be determined (i.e., both alleles missing), the SNP genotype was designated as M2. Coefficients of determination (R^2) calculated as squared correlations between estimated and true SNP genotypes (VanRaden et al., 2011), percentages of missing alleles (M1 and M2), and differenc-