Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Impact of Adding Foreign Genomic Information on Mexican Holstein Imputation

A. García-Ruiz*, *F. J. Ruiz-Lopez†, G. R. Wiggans‡, C. P. Van Tassell§ and H. H. Montaldo***,
*National Autonomous University of Mexico, D.F. México, †National Institute of Forestry, Agriculture and Livestock, Ajuchitlán, Querétaro, México ‡Animal Improvement Programs Laboratory and §Bovine Functional Genomics Laboratory, Agricultural Research Service, USDA, Beltsville, Maryland, USA

ABSTRACT: The impact of adding US and Canada genomic information to the imputation of Mexican Holstein genotypes was measured by comparing 3 scenarios: 1) 2,018 Mexican genotyped animals; 2) animals from scenario 1 plus 886 related North American animals; and 3) animals from scenario 1 and all North American genotyped animals (338,073). Four different chip densities were imputed to 45,195 markers using findhap software. Imputation success was measured by comparing the number of SNP half (HM), completely missing (CM) and conflicts. Imputation accuracy was improved when numbers of markers and genotyped animals were increased. The HM average was greater than the CM average for all scenarios. The largest number of different SNP filled (conflicts) was found between scenarios 1 and 3. The inclusion of genomic information of parents with daughters in the destination population improved accuracy imputation as did the inclusion of all available genotypes.
Keywords: Imputation; Genomics; Mexico-Holstein

Introduction

Increased reliability of genomic predictions is a highly sought-after result when using genomic information. To improve reliability calculations, increasing the number of genotyped animals is more important than using higher density panels for the reference population (VanRaden et al. (2010)). Unfortunately, genotyping is still expensive, especially in populations like the Mexican Holstein. Because high-density panels generally are more expensive than low-density panels, using the latter is an alternative that may lead to more genotyped animals. To use different marker densities, the information must be combined using imputation to predict missing genotypes of animals genotyped with lower density panels from genotype information of relatives or haplotypes of a population previously genotyped with higher density panels (Druet et al. (2010); VanRaden et al. (2010)). 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, thereby increasing the accuracy of genetic predictions and consequently the expected genetic improvement. Furthermore, reducing genotyping cost makes the technology more accessible to breeders (Berry et al. (2011); VanRaden et al. (2011)). For dairy cattle, reliability for genomic predictions using imputation varies according to trait and population size and structure. Reliability improvements of approximately 2 percentage points have been reported in simulation studies that included imputed genotypes compared with using a 50,000-marker subset (VanRaden et al. (2011)).

Currently genetic material from US and Canada dairy cattle is widely used around the world. Using genotypes for these animals for imputation could be a key point for the success of genomic evaluations and genetic progress in many countries. The objective of this study was to measure the impact of adding US and Canadian genomic information in the imputation of Mexican Holstein genotypes.

Materials and Methods

Scenarios. Three imputation scenarios were set up based on source and number of genotyped animals. For scenario 1, only genotypes of a local population of 2,018 genotyped Mexican Holsteins were included. For scenario 2, genotypes of animals in scenario 1 and genotypes of 866 North American Holsteins with genotyped Mexican daughters were included. For scenario 3, genotypes of animals in scenario 1 and all North American genotyped animals were included.

Genotypes. The Mexican Holstein population included a total of 1,978 cows and 40 sires from the Mexican conventional system were genotyped. For cows, 183 were genotyped with the Illumina BovineLD BeadChip v1.1 (6K), 277 with the GeneSeek Genomic Profiler LD BeadChip (9K), 686 with the Illumina BovineSNP50 BeadChip (50K), and 825 with the GeneSeek Genomic Profiler HD BeadChip (77K). All Mexican sires had 50K genotypes. From the North American database, 839 sires and 47 dams had genotyped progeny in Mexico and were included in scenario 2. All females in this group had 50K genotypes. Of the foreign sires with 50K genotypes, 533 were from the United States, 270 were from Canada, and 22 were from European countries; 14 sires had 77K genotypes: 10 from the United States and 4 from Canada. For scenario

3, a total of 338,073 North American animals were included in the analysis.

Pedigree information. Two different pedigree files were used in the analysis. One with 27,625 animals was used for scenarios 1 and 2; the other used for scenario 3 contained 938,662 animals.

Imputation. Missing genotypes were predicted by combining population and pedigree haplotypes with findhap software (VanRaden et al. (2011)). The imputation goal for the 3 scenarios was to fill in any missing genotypes from the 45,195 markers from 50K chip that is used in U.S. genomic evaluations. First, genotypes were recoded as 0 = BB, 1 = AB, 2 = AA, 3 = B and unknown maternal allele, 4 = A and unknown paternal allele, and 5 = both alleles unknown (VanRaden (2011)). Then the imputation for each scenario was performed, and the results were compared using SAS software (SAS (2009)). Imputation reliability, missing genotypes, and different filled markers or conflicts between scenarios were used for comparison. The imputation process is not always able to determine the genotype for a SNP. If neither parental contribution can be determined the genotype is designated as completely missing (genotype code of 5). If only one parental contribution can be determined, then the genotype is designated as half missing (code 3 or 4).

To test which imputed genotypes were more similar to true genotypes for scenarios 1 and 2, 10 groups of samples, each with 10 50K genotypes, were selected randomly. Non-6K and non-9Kmarkers were excluded, and the genotypes were filled through imputation. True and imputed genotypes were then compared for HM and CM markers. This validation was not performed for scenario 3 because of very small differences between results from scenarios 2 and 3 and because of large computational demands for scenario 3.

Results and Discussion

Accuracy. Imputation accuracies using only local genotypes (scenario 1) were high (96, 96, 99, and 99%, when imputing from 6K, 9K, 50K, and 77K chips, respectively). When information from North American Holsteins was added to Mexican genotypes (scenario 2), the imputation accuracy increased almost 1 percentage point for 6K and 9K chips and half a percentage point for the 77K chip. When all genotyped North American Holstein data were included (scenario 3) and compared with results for scenario 1, an increase of approximately 2 percentage points was observed for 6K and 9K chips and 1 percentage point for the 77K chip. As expected, no reliability increase was found for the 50K chip in any scenario because of the small number of SNP that actually were imputed. These results were consistent with those reported in other studies using the same (Wiggans et al. (2012)) or different (Zhang and Druet (2010); Johnston et al. (2011); Kathar et al. 2012)) imputation methods.

Missing alleles. Average number of CM and HM markers decreased from scenario 1 to 2 to 3 (Table 1). Numbers of animals with CM or HM markers did not necessarily follow this pattern. As expected, the 50K chip had fewer CM and HM markers, with the latter being substantially lower.

Table 1. Numbers of completely missing (CM) and half missing (HM) markers for 4 different densities of genotypes chips under 3 scenarios

Chip density	Scenario[1]	CM[2]	HM[2]
6,000	1	414 (183)	2,144 (183)
	2	23 (180)	1,426 (183)
	3	5 (26)	173 (183)
9,000	1	423 (277)	2,721 (277)
	2	18 (217)	1,615 (278)
	3	7 (25)	100 (277)
50,000	1	391 (726)	77 (726)
	2	14 (780)	94 (874)
	3	2 (196)	5 (690)
77,000	1	510 (820)	1,969 (820)
	2	33 (739)	1,005 (835)
	3	10 (153)	60 (815)

[1]Scenario 1: 2,018 Mexican genotyped animals; scenario 2: animals from scenario 1 plus 886 related North American animals; scenario 3: animals from scenario 1 and all North American genotyped animals (338,073).

[2]Number of animals within parentheses.

These results could be explained by the imputation process, because each chromosome first is divided into segments, all haplotypes are listed, and then all genotypes are matched with one from the haplotype list (VanRaden et al. (2011)). When the original genotype has more markers, haplotype matching is more precise and accuracy tends to be higher. However, the possibility of filling markers is less for genotypes with fewer markers. For this reason, more markers were filled as unknown (code 5) in the 77K genotypes, with an associated decrease in HM markers.

The number of HM markers was greater than for CM markers for all scenarios regardless of chip density, except for animals genotyped with the 50K chip in scenario 1. Number of HM markers followed the same pattern as CM markers between scenarios. However, between chips, the number of HM markers was greatest for the 9K chip followed by the 6K and 77K chips for scenarios 1 and 2. For scenario 3, the number of HM markers was for greatest for the 6K chip. The increased number of HM for the 50K chip for scenarios 2 and 3 could be the result of a larger haplotype list (because of more animals), which results in a better possibility of correct matches.

Conflicts. Conflicts were defined as SNP that were predicted differently in different scenarios. In the imputation process, the number of conflicts is of more concern than CM and HM numbers. Occurrences of CM and HM markers are possible to detect with the associated decrease in number of useful markers, but conflicts can lead

to estimation errors for future genomic studies (Weigel et al. (2010)). In this study, detected conflicts increased as the number of genotyped animals or the number of imputed SNP increased (Table 2). Conflict frequency never exceeded 3.8%, possibly because the 6K chip was the lowest density chip studied.

Table 2. Numbers of different imputed alleles (conflicts) for 3 scenarios

	Scenarios compared[1]	
Chip density	1 and 2	1 and 3
6,000	1,393	1,718
9,000	1,199	1,705
50,000	12	28
77,000	57	746

[1]Scenario 1: 2,018 Mexican genotyped animals; scenario 2: animals from scenario 1 plus 886 related North American animals; scenario 3: animals from scenario 1 and all North American genotyped animals (338,073).

Comparison of imputed and real genotypes. The average CM frequency for the 6K and 9K chips (Table 3) was similar (nearly 1%) regardless of scenario, whereas the average HM frequency was higher for scenario 1 compared with scenario 2. Genotypes that were imputed from the 9K chip had more HM markers on average than those imputed from the 6K chip for both scenarios, but the number of conflicts was highest for genotypes imputed from the 6K chip in scenario 1 (Table 3).

Table 3. Numbers of completely missing (CM) and half missing (HM) markers and conflicts for 2 different densities of genotyping chips when imputed and true genotypes were compared in 2 scenarios

Chip density	Scenario[1]	CM	HM	Conflicts
6,000	1	346	2,619	1,277
	2	538	1,860	1,022
9,000	1	346	2,906	1,181
	2	537	2,104	936

[1]Scenario 1: 2,018 Mexican genotyped animals; scenario 2: animals from scenario 1 plus 886 related North American animals.

When imputed and true genotypes were compared, the percentage of markers that were not useful [(CM + HM)/45,195, where 45,195 is the total number of 50K markers] was higher for 9K genotypes (7.2% for scenario 1 and 5.8% for scenario 2) than those for 6K genotypes (6.6 and 5.3%, respectively). However, 6K genotypes had a higher conflict rate (2.8% for scenario 1 and 2.3% for scenario 2) than did 9K genotypes (2.6 and 2.1%, respectively). For the 6K chip, similar error rates (2.7%) were reported when a panel of 6,000 markers was imputed to 45,836 markers in a reference population of 2,000 animals (Zhang and Druet (2010)). For the 9K chip, error rates were slightly lower than those reported (3.6 to 5.8%) for imputation of 8,680 markers in a reference population of 2,542 Jerseys (Weigel et al. (2010)). These results suggest that for low-density chips, imputation for 9K genotypes will be slightly more accurate than for 6K genotypes.

Conclusions

Imputation for Mexican Holsteins was affected by the size of the reference population and the number of markers in the original genotypes to be imputed. The inclusion of information from direct North American ancestors with genotypes in the imputation of Mexican Holstein genotypes increased imputation accuracy by half of what could be attained if genotype information from all North American Holsteins was included (approximately 1 versus 2 percentage points). Numbers of missing markers and imputation conflicts decreased when North American genotypes were included.

Literature Cited

Berry, D. P., and Kearney, J. F. (2011). Animal 5:1162–1169.

Browning, B. L., and Browning, S. R. (2011). Amer. J. Human Genet. 88:173–182.

Druet, T., Schrooten, C., and de Roos, A P. W. (2010). J. Dairy Sci. 93:5443–5454.

Hickey, J. M., Kinghorn, B. P., Tier, B. et al. (2012). Genet. Select. Evol. 44:9.

Johnston, J., Kistemaker, G., and Sullivan, P. G. (2011.) Interbull Bull. 44:25–33.

Khatkar, M. S., Moser, G., Hayes, B. J. et al. (2012). BMC Genomics 13:538.

SAS (2009). http://support.sas.com/documentation/cdl_main/index.html Accessed on February 25, 2014.

VanRaden, P. (2011). http://aipl.arsusda.gov/software/ findhap/ Accessed on February 25, 2014.

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R. et al. (2010). Interbull Bull. 42:113–117.

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R. et al. (2011). Genet. Sel. Evol. 43:10.

Weigel, K. A, de los Campos, G., Vazquez, A. I. et al. (2010). J. Dairy Sci. 93:5423–5435.

Wiggans, G. R., Cooper, T. A, VanRaden, P. M. et al. (2012). J. Dairy Sci. 95:1552–1558.

Zhang, Z., and Druet, T. (2010). J. Dairy Sci. 93:5487–5494.