J. Dairy Sci. 100:1–12
https://doi.org/10.3168/jds.2017-12732

Genotype imputation in a tropical crossbred dairy cattle population

Gerson A. Oliveira Júnior,* Tatiane C. S. Chud,† Ricardo V. Ventura,‡§ Dorian J. Garrick,# John B. Cole,‖ Danísio P. Munari,† José B. S. Ferraz,* Erik Mullart,¶ Sue DeNise, Shannon Smith,** and Marcos Vinícius G. B. da Silva††[1]**

*Departamento de Medicina Veterinária, Universidade de São Paulo (USP), Faculdade de Zootecnia e Engenharia de Alimentos, Pirassununga, SP, 13635-900, Brazil
†Departamento de Ciências Exatas, Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, SP, 14884-900, Brazil
‡Beef Improvement Opportunities, Guelph, ON N1K1E5, Canada
§Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, ON N1G2W1, Canada
#Department of Animal Science, Iowa State University, Ames 50011-3150
‖Animal Genomics and Improvement Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD, 20705-2350
¶CRV Holding B.V., Arnhem, 454, the Netherlands
**Zoetis, Kalamazoo, MI 49007
††Embrapa Dairy Cattle, Brazilian Corporation of Agricultural Research, Juiz de Fora, MG, 36038-330, Brazil

ABSTRACT

The objective of this study was to investigate different strategies for genotype imputation in a population of crossbred Girolando (Gyr × Holstein) dairy cattle. The data set consisted of 478 Girolando, 583 Gyr, and 1,198 Holstein sires genotyped at high density with the Illumina BovineHD (Illumina, San Diego, CA) panel, which includes ∼777K markers. The accuracy of imputation from low (20K) and medium densities (50K and 70K) to the HD panel density and from low to 50K density were investigated. Seven scenarios using different reference populations (RPop) considering Girolando, Gyr, and Holstein breeds separately or combinations of animals of these breeds were tested for imputing genotypes of 166 randomly chosen Girolando animals. The population genotype imputation were performed using FImpute. Imputation accuracy was measured as the correlation between observed and imputed genotypes (CORR) and also as the proportion of genotypes that were imputed correctly (CR). This is the first paper on imputation accuracy in a Girolando population. The sample-specific imputation accuracies ranged from 0.38 to 0.97 (CORR) and from 0.49 to 0.96 (CR) imputing from low and medium densities to HD, and 0.41 to 0.95 (CORR) and from 0.50 to 0.94 (CR) for imputation from 20K to 50K. The $CORR_{anim}$ exceeded 0.96 (for 50K and 70K panels) when only Girolando animals were included in RPop (S1). We found smaller $CORR_{anim}$ when Gyr (S2) was used instead of Holstein (S3) as RPop. The same behavior was observed between S4 (Gyr + Girolando) and S5 (Holstein + Girolando) because the target animals were more related to the Holstein population than to the Gyr population. The highest imputation accuracies were observed for scenarios including Girolando animals in the reference population, whereas using only Gyr animals resulted in low imputation accuracies, suggesting that the haplotypes segregating in the Girolando population had a greater effect on accuracy than the purebred haplotypes. All chromosomes had similar imputation accuracies ($CORR_{snp}$) within each scenario. Crossbred animals (Girolando) must be included in the reference population to provide the best imputation accuracies.

Key words: impute, single nucleotide polymorphism, genotype

INTRODUCTION

In tropical and subtropical countries, such as Brazil, crossbred animals mainly result from matings between taurus (*Bos taurus*) and zebu (*Bos indicus*) animals. Such crosses have been widely used by farmers and breeders in both beef and dairy cattle industries. This practice, in dairy cattle, exploits complementarity, combining higher milk production present in taurines with heat tolerance and parasite resistance present in indicine breeds. The Girolando is an example of a crossbred dairy breed, resulting from crossbreeding between Holstein and Gyr cattle, with genetic composition ranging from 1/4 to 7/8 Holstein. In Brazil, which is one of the largest milk producing countries, 80% of milk production comes from crossbred cattle, mostly of the Girolando breed (Cole and da Silva, 2016), with an estimated population close to 10 million animals.

Received February 14, 2017.
Accepted August 16, 2017.
[1]Corresponding author: marcos.vb.silva@embrapa.br

Genomic evaluation has been successfully implemented, particularly in Holstein breed, and has allowed substantial increases in rates of genetic gain (Hayes et al., 2009; Wiggans et al., 2011; Olson et al., 2012; García-Ruiz et al., 2016). However, to maximize the reliability of genomic breeding values it is necessary to have many individuals in the reference population genotyped for thousands of SNP markers (Goddard, 2009). Most crossbred dairy populations (e.g., Girolando) have only a small number of progeny-tested bulls available to construct the reference population (Cole and da Silva, 2016); consequently, these populations have low reliability genomic breeding values (Thomasen et al., 2013). An alternative to increasing the number of animals in the reference population of the target breed is to combine data sets from related purebred and multibred populations (Lund et al., 2014).

Genotype imputation is a well-established statistical technique for using known marker information to infer unknown marker information, such as imputing low-density cow genotypes up to medium or high density (VanRaden et al., 2011). Genotype imputation has been used to reduce costs of genotyping and to combine data sets from different breeds and chip densities (Howie et al., 2011; Khatkar et al., 2012; Larmer et al., 2014). Low-density panels may be imputed to higher density using information of haplotype segments from densely genotyped animals in the reference population (Ventura et al., 2014). Several studies using different strategies and methodologies for genomic imputation have shown satisfactory results in crossbred cattle (Ventura et al., 2014; Chud et al., 2015; Jattawa et al., 2016) and in other species such as swine (Cleveland and Hickey, 2013; Xiang et al., 2015) and sheep (Bolormaa et al., 2015; Ventura et al., 2016).

Several factors can influence imputation accuracy, such as the number of animals in the reference population (Khatkar et al., 2012; Ventura et al., 2014), allele frequency of the imputed SNP (van Binsbergen et al., 2014; Boison et al., 2015), the SNP density on the low and high panel (Carvalheiro et al., 2014; Judge et al., 2016), relatedness between individuals in the reference and target populations (Boison et al., 2015; Ventura et al., 2016), and imputation methods used (Chud et al., 2015; Ventura et al., 2016). According to Moghaddar et al. (2015), imputation accuracy increases for both purebred and crossbred animals when breed-specific haplotypes are available in the reference population.

The implementation of some genomic methods, such as genotype imputation, are still challenging in crossbred populations. This is the first imputation study considering a Girolando population. The aim of this research was to quantify the accuracies of different strategies for genotype imputation in a crossbred Girolando dairy cattle population that is important in tropical dairy production.

MATERIALS AND METHODS

Data Set

All Gyr and Girolando data used for this study were provided by dairy breeding programs in Brazil, where the Brazilian Corporation of Agricultural Research (Embrapa Dairy Cattle), located in Juiz de Fora, MG, Brazil, is the institution responsible for genetic evaluations. The Holstein genotype data were provided by Zoetis (Kalamazoo, MI) and CRV BV, Arnhem, the Netherlands.

The database consisted of 478 Girolando, 583 Gyr, and 1,198 Holstein sires genotyped using the Illumina BovineHD BeadChip (Illumina, San Diego, CA) panel (**HD**) comprising 777,962 markers distributed throughout the genome. The pedigree information of these genotyped Girolando animals consisted of 5,404 animals, including 970 sires and 2,544 dams. Of these animals, 2,924 had information on both parents and 288 animals had at least one known parent. The genomic relationship matrix (**G**) was estimated for each breed separately according to the method of VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{MM'}}{2\Sigma p_i \left(1 - p_i\right)},$$

in which **M** is the incidence matrix of markers whose elements in the ith column are 0–2× p_i, 1–2× p_i, and 2–2× p_i for genotypes AA, AB, and BB, respectively; $\mathbf{M}'$ is the transpose of the incidence matrix; summation is over the number of marker loci; and p_i is the within-breed frequency of allele B for the ith marker. The average of genomic inbreeding coefficient (one minus the diagonals of **G** matrix) was 0.04% for Girolando, 0.87% for Gyr, and 0.80% for Holstein.

The Girolando animals in this study had an average composition of 0.34 Gyr and 0.66 Holstein ancestry as estimated by Admixture v1.3. (Alexander et al., 2009). Principal component analysis (**PCA**) were carried out to describe the purebreds (Gyr and Holstein) and Girolando breed. The PCA analysis was calculated from the genomic matrix using the function -pca from PLINK v1.9 software (Chang et al., 2015).

Low- and Medium-Density Panels and Imputation Scenarios

Low- and medium-density panels were simulated from the HD genotypes by selecting markers present on