J. Dairy Sci. 100:9623–9634
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

Table 1. Number of animals and SNP markers in the reference population for each scenario used to impute the validation population of 166 Girolando animals

	Reference animals[1]			
Scenario	Girolando	Gyr	Holstein	SNP
S1a	300	0	0	672,348
S2a	0	557	0	584,674
S2b	0	H300ped	0	728,575
S2c	0	L300ped	0	728,323
S3a	0	0	805	658,427
S3b	0	0	H300ped	642,074
S3c	0	0	L300ped	650,392
S4a	300	557	0	550,882
S4b	300	H300ped	0	615,931
S4c	300	L300ped	0	616,645
S5a	300	0	805	718,267
S5b	300	0	H300ped	713,608
S5c	300	0	L300ped	719,807
S6a	0	557	805	351,088
S6b	0	557	H300ped	379,449
S6c	0	557	L300ped	373,947
S6d	0	H300ped	805	413,494
S6e	0	L300ped	805	417,192
S6f	0	H300ped	H300ped	416,257
S6g	0	L300ped	L300ped	414,266
S7a	300	557	805	378,999
S7b	300	557	H300ped	407,551
S7c	300	557	L300ped	409,992
S7d	300	H300ped	805	442,614
S7e	300	L300ped	805	440,836
S7f	300	H300ped	H300ped	471,617
S7g	300	L300ped	L300ped	470,888
S7h	300	L300ped	H300ped	475,743
S7i	300	H300ped	L300ped	462,533

[1]H300ped = sires with highest numbers of progeny in the pedigree file; L300ped = sires with lowest numbers of progeny in the pedigree file.

the following commercial genotyping platforms: Zoetis Custom SNP chip ZL2 (20K), Illumina BovineSNP50 BeadChip v2 (50K), and Zoetis Custom SNP chip ZM2 (70K).

We considered 2 strategies for imputation: (1) from 20K, 50K, or 70K panels to the HD panel, or (2) from the 20K to the 50K panel. Several scenarios were investigated using different reference populations (**RPop**) for imputation. First, genotyped animals from the studied breeds were considered separately (**S1**, Girolando; **S2**, Gyr; and **S3**, Holstein), and second, 2- or 3-way combinations of animals of these breeds were formed (**S4**, Girolando + Gyr; **S5**, Girolando + Holstein; **S6**, Gyr + Holstein; and S7, Girolando + Gyr + Holstein; Table 1). The target population (**TPop**) always comprised the same 166 randomly chosen animals among the 478 genotyped Girolando animals that were not part of the reference population. Different sets of randomly chosen animals in Tpop were tested in different scenarios, with no effect on the results.

Variable numbers of animals were tested in the purebred RPop groups, considering all available data or smaller subsets of animals from each breed. To investigate the importance of the relationships between animals in the imputation analyses, 2 groups of sires were formed, keeping in RPop the 300 with most (**H300ped**) or fewest (**L300ped**) progeny in the pedigree file. To verify the effect of the relationship between individuals from reference and target population we verified the average relationship between each animal and the 5 (top 5) and 10 (top 10) individuals most related in Rpop and Tpop.

Genotype Quality Control

Genotype quality control and imputation were performed using only SNP from autosomal chromosomes with known positions on the UMD 3.1 reference genome assembly (Zimin et al., 2009). The filters were applied in each scenario discarding markers with call rates lower than 0.90, minor allele frequency (**MAF**) <0.001, or when the observed percentage of heterozygous markers differed from expected (Hardy-Weinberg equilibrium) by >15%. Animals with genotype call rates lower than

0.90 were removed from the data set. The final number of available markers after applying these filters varied according to each scenario (Table 1).

Genotype Imputation and Accuracy

FImpute (Sargolzaei et al., 2014) was used for genotype imputation based on results of performance comparisons reported in the literature (Ventura et al., 2014; Chud et al., 2015; Jattawa et al., 2016). It uses a population-based methodology assuming that all individuals have some degree of relationship to one another. Long to short sliding windows are built to capture information from close to far relatives to help identify shared haplotypes for accurate phasing and imputation using an iterative approach as described in Sargolzaei et al. (2014).

Sample-specific imputation accuracies were calculated via Pearson correlation between observed and imputed genotypes ($\mathbf{CORR_{anim}}$), and also as the imputation concordance rate ($\mathbf{CR_{anim}}$), which represents the proportion of correctly imputed genotypes. The $\text{CORR}_{\text{anim}}$ has been reported (Hickey et al., 2012; van Binsbergen et al., 2014; Boison et al., 2015) to have the advantage of not being dependent on allele frequency, and is defined as follows:

$$\text{CORR}_{\text{anim}} = \frac{\sum_{j=1}^{L_k}\left(g_{ij}-\overline{g}\right)\left(\hat{g}-\overline{\hat{g}}\right)}{\sqrt{\sum_{j=1}^{L_k}\left(g_{ij}-\overline{g}\right)^2\sum_{j=1}^{L_k}\left(\hat{g}_{ij}-\overline{\hat{g}}\right)^2}},$$

where L is the total number of markers being imputed; g_{ij} and $\hat{g}_{ij}$ are the observed and imputed genotypes for SNP j of individual i; $\overline{g}_{ij}$ and $\overline{\hat{g}}_{ij}$ are the average values of observed and imputed genotypes, respectively.

For SNP-specific imputation, accuracies were defined based on the same criteria (CORR_{snp}, CR_{snp}), but summing over N_k (total number of animals for marker k) rather than L_k (the number of filtered markers being imputed). Markers were classified into 4 groups of MAF to check the effect of MAF in SNP imputation accuracy, being a (MAF ≤ 0.05), b ($0.05 <$ MAF ≥ 0.10), c ($0.10 <$ MAF ≥ 0.20), and d (MAF > 0.20).

RESULTS AND DISCUSSION

Genotype Imputation Accuracy

The PCA analysis showed that the 166 crossbred animals included in the TPop were randomly chosen from the Girolando population (Figure 1). Similarly, the subsets of H300ped and L300ped animals were also randomly distributed in the population in terms of the first 2 eigenvectors.

The $\text{CORR}_{\text{anim}}$ and CR_{anim} ranged from 0.38 to 0.97 and 0.49 to 0.96 for imputation from 20K, 50K, and 70K to the HD panel (Figure 2 and Table 2; Supplemental Figure S1 and Table S1, https://doi.org/10.3168/jds.2017-12732); from 0.41 to 0.95 and from 0.50 to 0.94 from 20K to 50K (Supplemental Table S2, https://doi.org/10.3168/jds.2017-12732), respectively. The $\text{CORR}_{\text{anim}}$ exceeded 0.96 (for 50K and 70K panels) when only Girolando animals were included in RPop (S1). Slightly larger values were observed in other scenarios that also included crossbreds in the RPop (S4, S5, and S7), whereas lower values were achieved in other scenarios, with a difference up to 0.55 when comparing S2 and S1.

The imputation methodology used by FImpute constructs haplotype segments that are as large as possible and iteratively moves to smaller ones if no consistent haplotypes are found in RPop. When animals are separated from their common ancestor by many generations conserved genomic blocks can consist of very short haplotypes. This may explain the smaller $\text{CORR}_{\text{anim}}$ when Gyr (S2) was used instead of Holstein (S3) as RPop. The same behavior was observed between S4 (Gyr + Girolando) and S5 (Holstein + Girolando), with the latter having higher $\text{CORR}_{\text{anim}}$ because the TPop animals were more related to the Holstein population.

Given this small increase in accuracy, the results suggest that using only Girolando animals in the reference population (S1) was able to account for the majority of the haplotypes presents in the TPop. Piccoli et al. (2014) did not observe significant differences in imputation accuracies when Nellore animals were included in RPop to impute a set of crossbred Braford (Zebu × Hereford). Berry et al. (2014) reported no gains in imputation accuracies when multiple breeds were included in RPop and concluded that imputation accuracy was improved when RPop only included animals from the breed to be imputed. The lowest imputation accuracies were observed ($\text{CORR}_{\text{anim}}$: 0.42–0.51; CR_{anim}: 0.51–0.58) when only Gyr animals were included in RPop (S2), suggesting not surprisingly that animals from RPop and TPop shared a small number of recent haplotypes. Ventura et al. (2016) reported similar results in an investigation of sheep breeds when a different breed was used in the TPop, imputation accuracy was similar to results from a SNP set including only the most frequent marker alleles. Given these results, more animals should be included in the RPop set to better ensure coverage of common haplotypes.

Following Carvalheiro et al. (2014), the $\text{CORR}_{\text{anim}}$ values were higher than the corresponding values in

CR_{anim}, except when only Gyr animals comprised RPop (S2), where the difference between CR_{anim} and $CORR_{anim}$ reached 0.09 when imputing from 20K to the HD panel. This difference in imputation accuracy difference in S2 may be explained by the higher number of markers with low MAF. As reported by Hickey et al. (2012), the probability of being homozygous for the common allele increases for a marker with very low MAF, which consequently produces, in general, an increased CR. The opposite behavior is expected for CORR, where the imputation accuracy is lower for markers with low MAF.

As previously reported (Hickey et al., 2012; Pausch et al., 2013; Piccoli et al., 2014), the imputation accuracy increased as the number of genotypes that were masked decreased. More genotyped SNP around missing markers are available when imputing from medium-density panels (50K and 70K), and small haplotype fragments that are conserved over generations and may be easily detected compared with lower-density panels (Weng et al., 2013; Larmer et al., 2014). Increasing the number of genotyped markers around missing SNP provide more information that could be used to construct haplotypes and to infer the missing SNP, and therefore result in better imputation accuracies.

Druet et al. (2010) argued that the benefits of using denser panels in TPop were more evident when compared with the low-density panels. Likewise, because the 50K and 70K share close to 89% and similar distributions across chromosomes (Table 3), only small differences in $CORR_{anim}$ were observed between them. Boison et al. (2014) showed that parent-offspring pairs shared genomic regions that can be precisely identified, achieving accuracies above 0.90. This suggests that, depending on the composition of RPop, genotyping animals using the 50K might be enough to impute to an HD panel. However, when Girolando animals were included in RPop (S1, S4, S5, and S7), imputation from 20K to HD was also quite accurate ($CORR_{anim} > 0.92$), and the use of a lower-density panel will permit the ge-

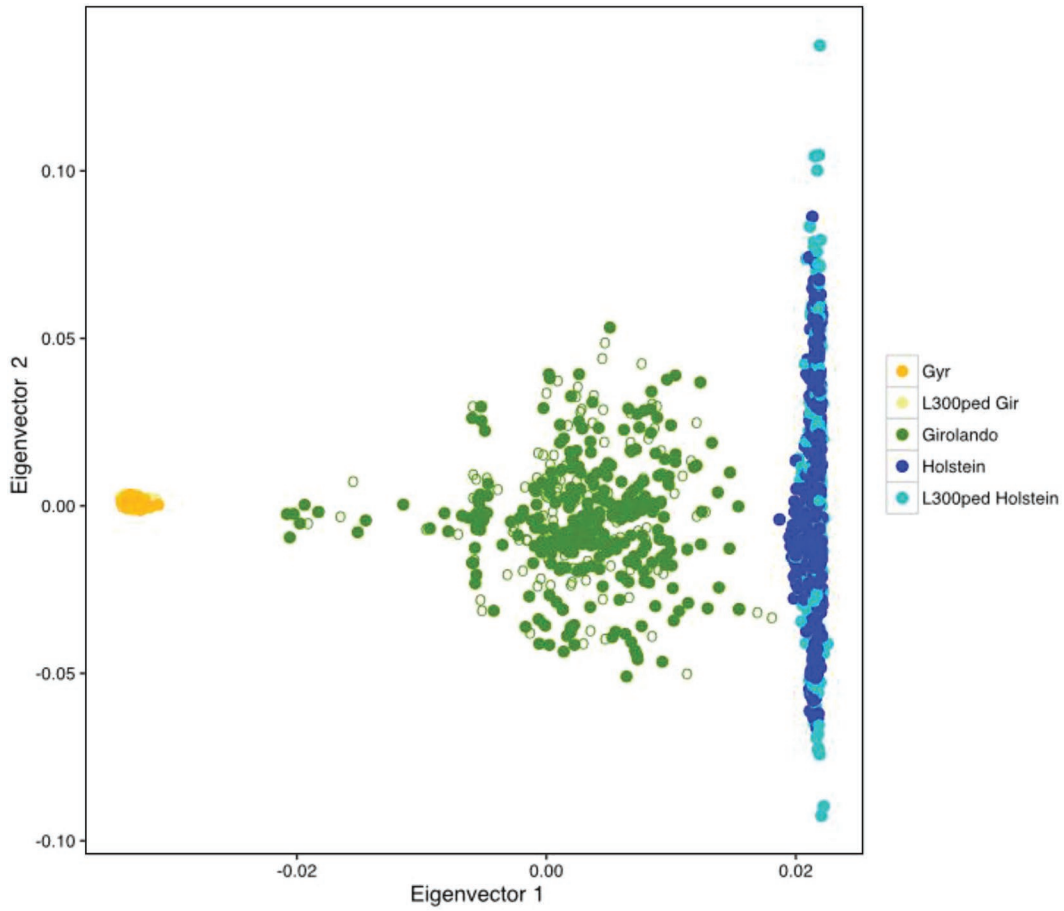

Figure 1. Eigenvector 1 (30.03%) and eigenvector 2 (0.98%) of the genomic relationship matrix. Open green dots represented the animals included in the imputation validation population (166 Girolando). The lighter colors represent the 300 animals with lower (L300ped) numbers of progeny in the pedigree file. Color version available online.

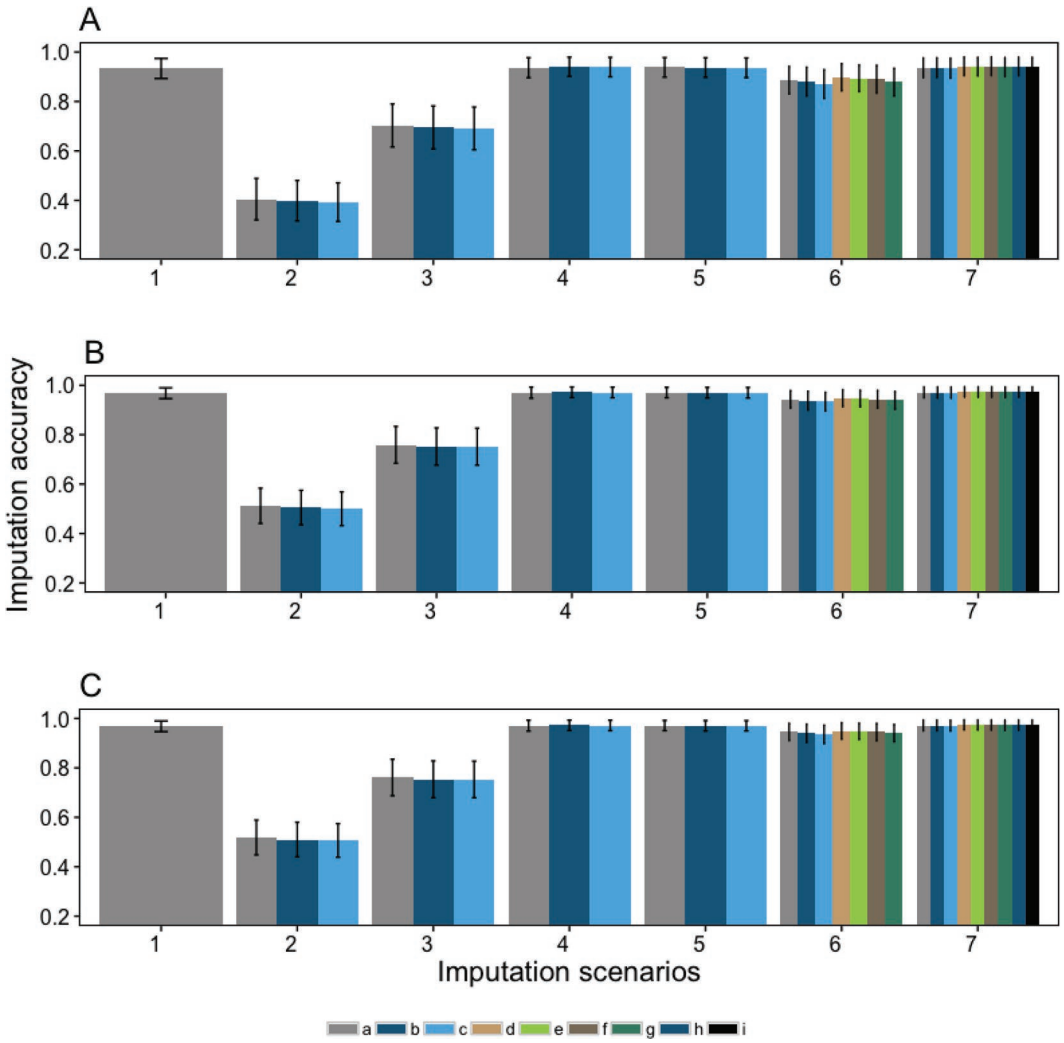

Figure 2. Average and SD of animal imputation accuracy ($CORR_{anim}$) of genotype imputation from 20K (A), 50K (B), and 70K (C) to the Illumina BovineHD (HD; Illumina, San Diego, CA) panel, considering all imputation scenarios (1–7). Color version available online.

Table 2. Average animal imputation accuracy from low (20K) and medium densities (50 and 70K) to Illumina BovineHD (HD; Illumina, San Diego, CA) panel, considering different scenarios for the reference population and 166 Girolando animals in the main imputation population[1]

	20K		50K		70K	
Scenario	CORR	CR	CORR	CR	CORR	CR
S1a	0.93	0.92	0.96	0.96	0.96	0.96
S2a	0.42	0.51	0.51	0.59	0.51	0.59
S3a	0.70	0.70	0.75	0.74	0.75	0.74
S4a	0.93	0.92	0.97	0.96	0.97	0.96
S5a	0.93	0.93	0.97	0.96	0.97	0.96
S6a	0.88	0.86	0.94	0.93	0.94	0.93
S7a	0.93	0.92	0.97	0.96	0.97	0.96

[1]CORR = correlation between observed and imputed genotypes; CR = proportion of genotypes that were imputed correctly.

Table 3. Number of SNP markers on autosome chromosomes shared between commercial panels Zoetis Custom SNP chip ZL2 (20K; Zoetis, Kalamazoo, MI), Illumina BovineSNP50 v2 (50K; Illumina, San Diego, CA), and Zoetis Custom SNP chip ZM2 (70K), and Illumina BovineHD (HD)

Item	20K	50K	70K	HD
20K	16,706	10,790	15,075	15,643
50K	—	52,856	46,973	47,787
70K	—	—	57,901	52,869
HD	—	—	—	735,239

notyping of more animals. Similar results were reported by Carvalheiro et al. (2014) and Boison et al. (2015) in Nellore and Gyr animals, respectively.

SNP Imputation Accuracy

All chromosomes had similar imputation accuracies ($CORR_{snp}$) within each scenario (Figure 3; Supplemental Figure S2, https://doi.org/10.3168/jds.2017-12732), in agreement with Jattawa et al. (2016) using FImpute in a crossbred dairy population. *Bos taurus* autosome 14 had the least accurate imputation for S1 (0.91) and S6 (0.90), and BTA15 and BTA27 for S2 (0.30) and S3 (0.64), respectively, when imputing from 50K to HD. Chud et al. (2015) also reported BTA27 as the least accurate from 50K and 80K to HD in a Canchim (Charolais × Zebu breeds) population. Piccoli et al. (2014) and Ventura et al. (2016) argued that due to the complexity of imputation at chromosome ends, the imputation accuracy is normally poorer for shorter chromosomes.

Imputation accuracy increased with increasing MAF (Figure 4, Supplemental Figure S3; https://doi.org/10.3168/jds.2017-12732) as suggested by Chud et al. (2015). The number of markers with MAF less than 0.05 (class *a*) was bigger for S2, where about 23% of markers fell in that class. This inaccuracy when imputing low MAF SNP (Hickey et al., 2012; Sargolzaei et al., 2014) explains the poorer results for this scenario when compared with the others. For Sargolzaei et al. (2014), rare alleles (MAF <0.05) might be recent mutations and are easily recognized after identifying long haplotype blocks. Wray (2005) discuss that the imputation of rare alleles is important because causal mutations may be in linkage disequilibrium with them.

Effect of Relatedness on Imputation Accuracy

The $CORR_{anim}$ for the 2 first scenarios (single purebreds in RPop) were clearly influenced by the relatedness between breeds (Figure 5, Supplemental Figure S4; https://doi.org/10.3168/jds.2017-12732). Better $CORR_{anim}$ were achieved for animals with a higher

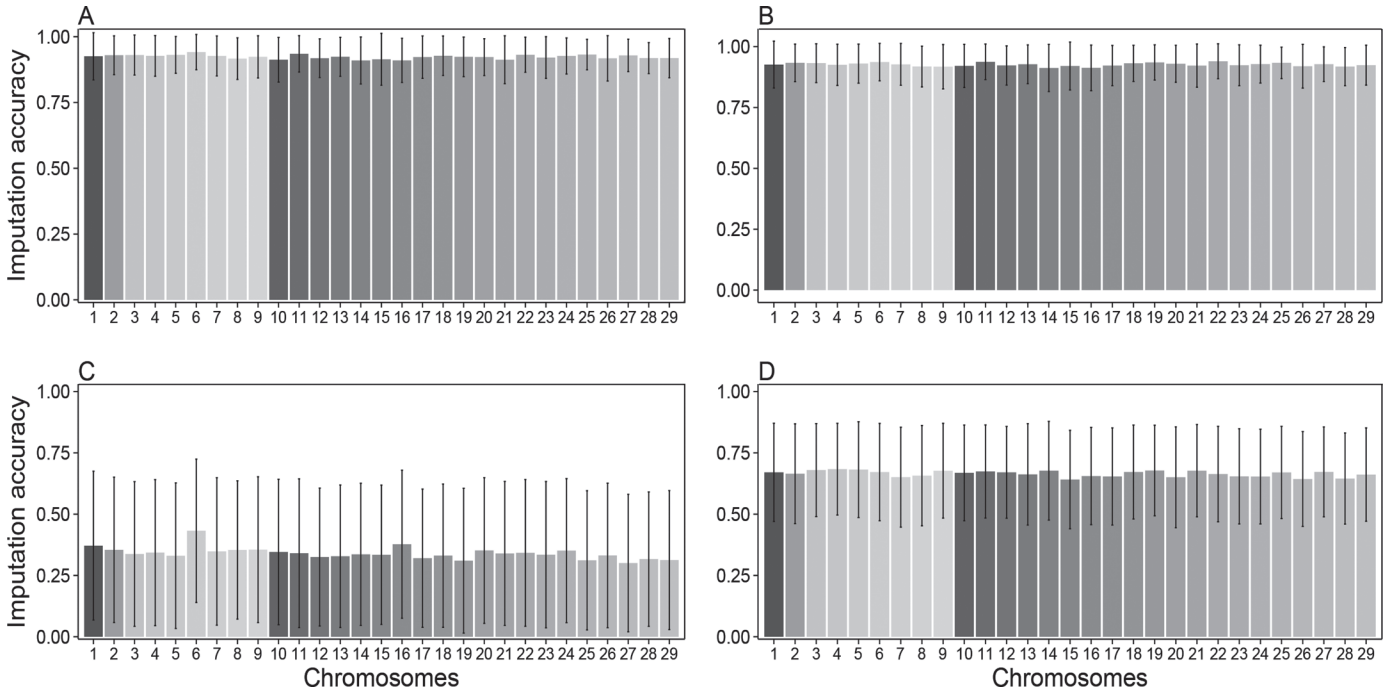

Figure 3. Average SNP-specific imputation accuracy ($CORR_{snp}$) and SD (vertical lines) by chromosome from 50K to Illumina BovineHD (HD; Illumina, San Diego, CA) panel, considering the scenarios that included animals from Gyr + Girolando + Holstein breeds (A), only Girolando (B), Gyr (C), or Holstein (D) animals.

proportion of the breed included in RPop. Moghaddar et al. (2015) also reported better imputation accuracy when animals in RPop were genetically more related to the TPop. The $CORR_{anim}$ was independent of the genomic ancestry of the purebreds when the crossbreds were included in the RPop (Figure 5A, 5B). However, the accuracy was dependent on the animal composition when just Gyr (Figure 5C) or Holstein (Figure 5D) animals were included in the RPop used.

Ventura et al. (2014) argued that the inclusion of ancestors in RPop is necessary for accurate imputation, reporting a 3% gain in overall imputation accuracy from including key relatives in RPop. Bolormaa et al. (2015) also reported better imputation accuracy when closely related animals were included in RPop. Likewise, in this study, the use of H300ped over L300ped information as RPop produced favorable gains in imputation accuracy. In all scenarios, the H300ped animals performed better than L300ped as RPop, with the largest difference between scenarios S2b and S2c (0.03), S6f and S6g (0.01), and S6b and S6c (0.01) when imputing from 20K to HD. The variation was even larger for scenarios 2b and 2c (0.04) when imputing from 20K to 50K. Scenarios that included Girolando animals in RPop showed only slight differences, with the smallest effect observed between scenarios 5b and 5c (less than 0.001).

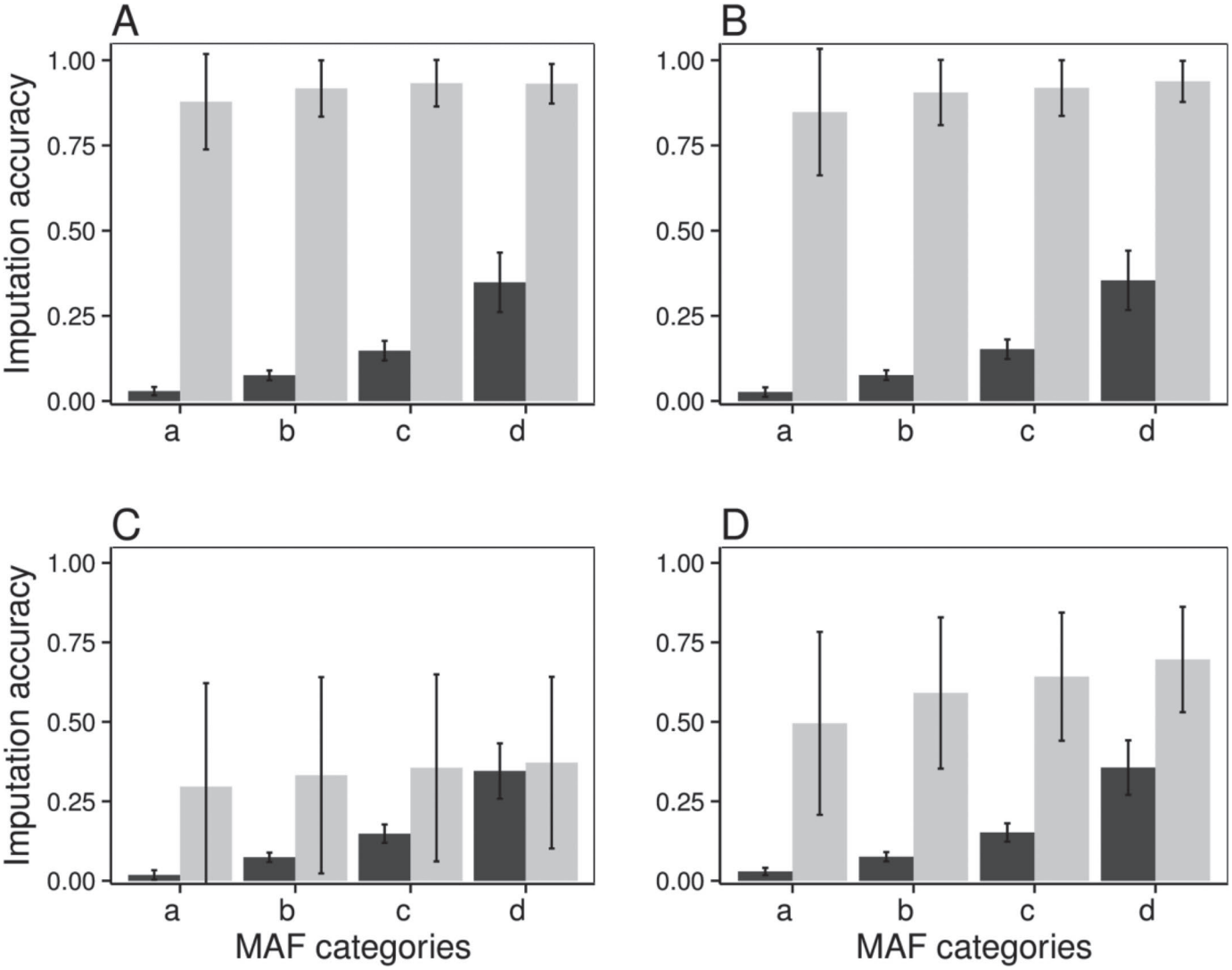

Figure 4. Average SNP-specific imputation accuracy ($CORR_{SNP}$, gray) and average minor allele frequency (MAF; black) with respective SD (vertical lines) for different classes of MAF when imputing from 50K to Illumina BovineHD (HD; Illumina, San Diego, CA) using reference population of Gyr + Girolando + Holstein breeds (A), Girolando (B), Gyr (C), or Holstein (D). The MAF classes were *a* (MAF $\leq$ 0.05), *b* (0.05 $<$ MAF $\leq$ 0.10), *c* (0.10 $<$ MAF $\leq$ 0.20), and *d* (MAF $>$ 0.20).

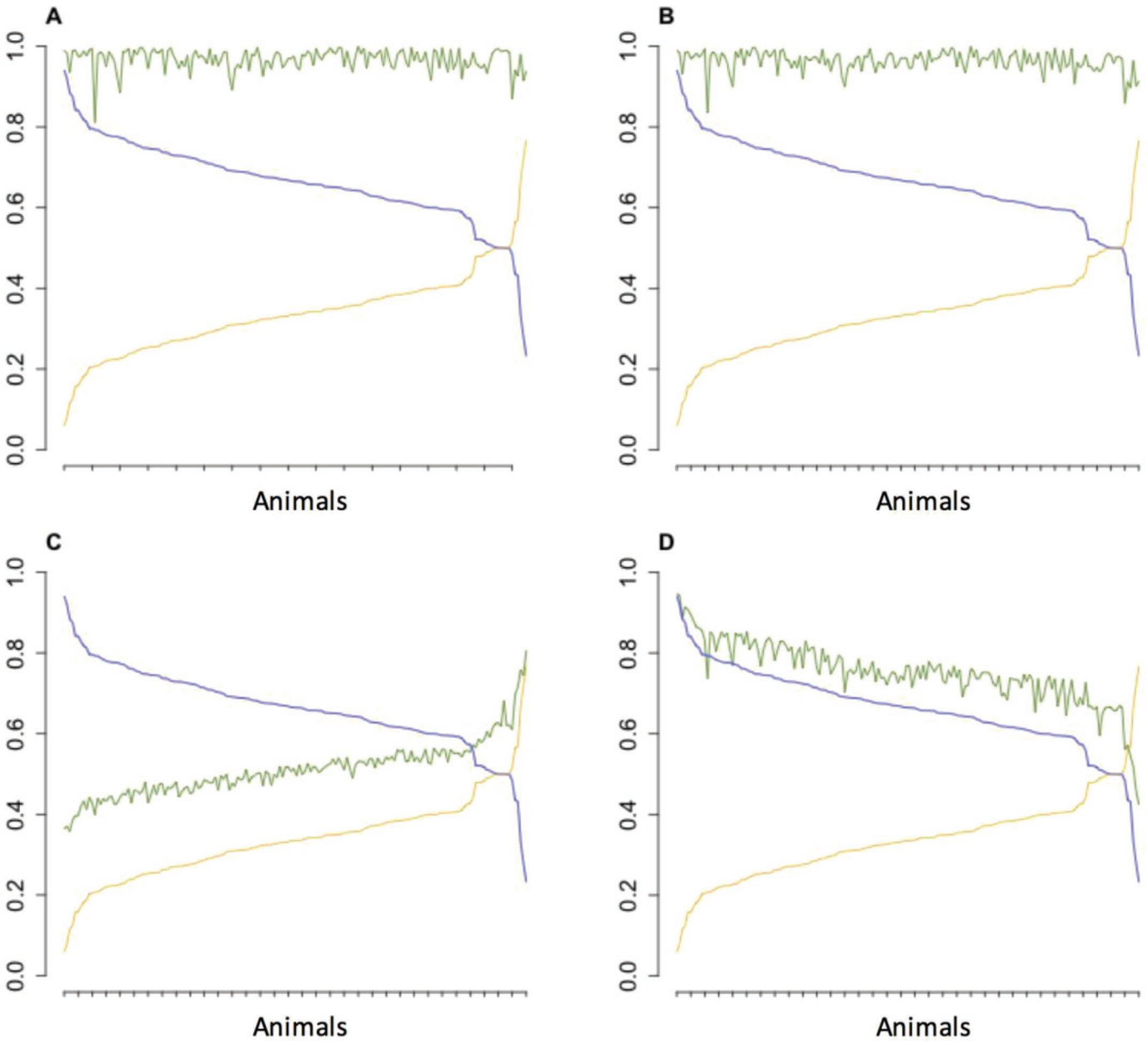

Figure 5. The breed composition of the 166 Girolando animals included in the imputed population (x-axis). The dark gray (blue) and light gray (yellow) lines represent the proportion of Holstein and Gyr, respectively, whereas the mid gray (green) line is the imputation accuracy ($CORR_{anim}$) when considered as reference population animal from Gyr + Girolando + Holstein breeds (A), only Girolando (B), Gyr (C), or Holstein (D) animals from 50K to Illumina BovineHD (HD; Illumina, San Diego, CA). Color version available online.

Similarly, scenarios with a stronger relationship between animals in RPop and TPop were less influenced by the top 5 and top 10 measurement (Figure 6A, 6B, 6D and Supplemental Figure 5A, 5B, 5D; https://doi.org/10.3168/jds.2017-12732). These results reinforce the fact that closely related animals share longer haplotype segments, which are used in the inference of missing markers. Boison et al. (2015) evaluated imputation accuracy for the Gyr population, reporting that when the average relationship between each imputed animal and the top 5 animals in the reference population was lower than 0.10, it could be an indication of lower imputation accuracy (less than 0.90). For Bolormaa et al. (2015) and Ventura et al. (2016), the relationship measurement demonstrated that the imputation accuracy is highly dependent on the genetic distance between

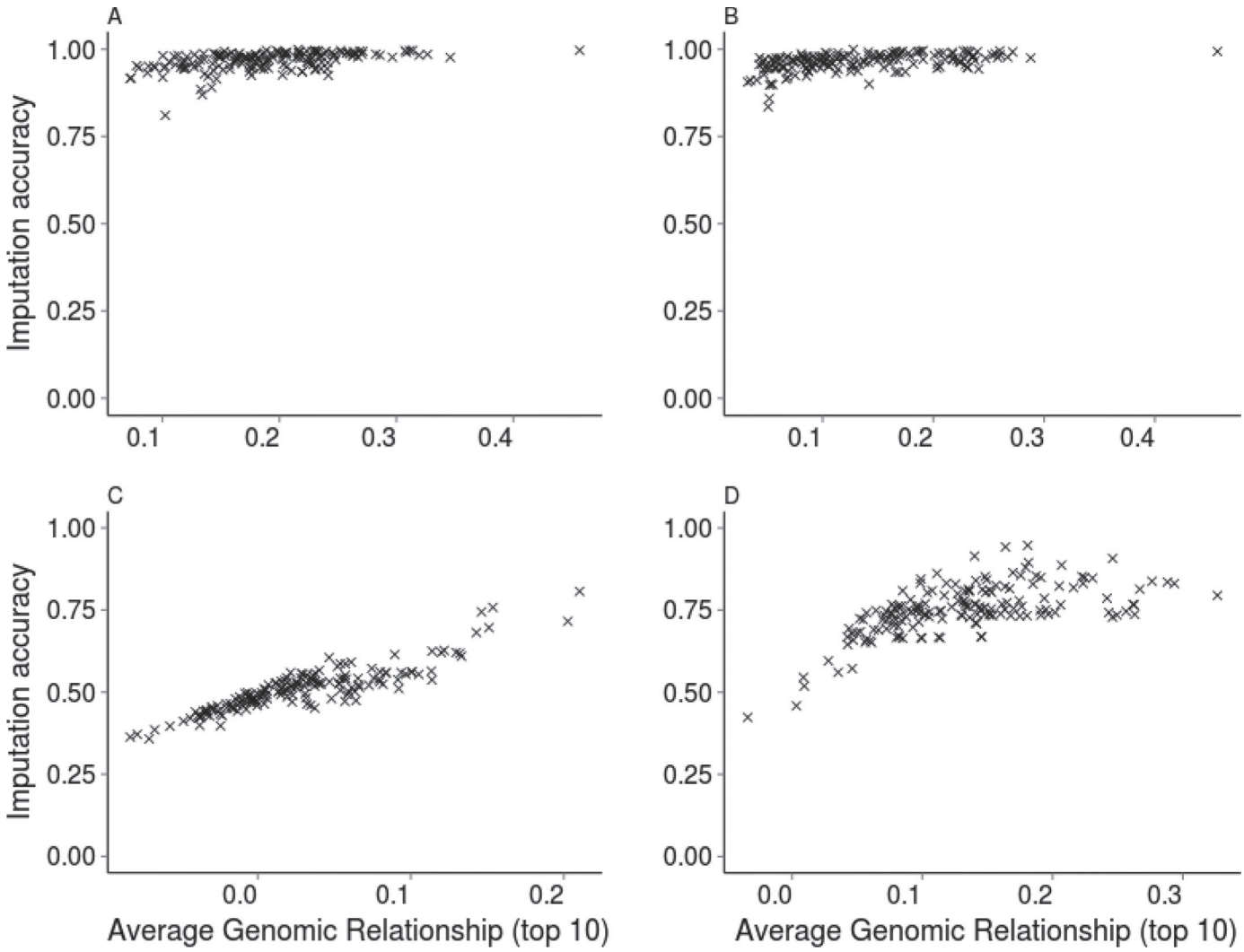

Figure 6. Plot of imputation accuracy ($CORR_{anim}$) when imputing from 50K to Illumina BovineHD (HD; Illumina, San Diego, CA) as a function of the genomic average relationship between each imputed animal and the 10 most related (top 10) animals in the reference sets (A = Gyr + Girolando + Holstein breeds; B = Girolando; C = Gyr; D = Holstein). $CORR_{anim}$ = correlation between true and imputed genotypes.

the animals in the TPop and RPop. In other words, genetically closer animals in the reference and imputation population produce higher imputation accuracies.

CONCLUSIONS

This was the first study to evaluate the imputation accuracy in a Girolando population, and these results may provide information to assist future studies involving genomic data in crossbred animals. 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. Crossbred animals (Girolando) must be included in the reference population to provide the best imputation accuracies. The results obtained in this work provide information to more cost effectively implement genomic selection and will assist future studies involving genomic analysis in crossbred animals.

ACKNOWLEDGMENTS

The authors declare that they have no competing interests. The authors acknowledge the Brazilian Corporation of Agricultural Research (Embrapa), Universidade Estadual Paulista–Faculdade de Ciências Agrárias e Veterinárias, Iowa State University, Universidade de São Paulo–Faculdade de Zootecnia e Engenharia de Alimentos, Zoetis, CRV, and the Agricultural Research

Service of the USDA. We are also grateful to Mehdi Sargolzaei (L'Alliance Boviteq, Saint-Hyacinthe, QC, Canada; and Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, Ontario, Canada) for sharing the FImpute software. GAOJ and TCSC were Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) fellows (grants no. 2013/12097-9 and 2015/08939-0). DPM, JBF, and MVGBS are Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) fellows. Marcos V. G. B. Silva was supported by the Embrapa (Brazil) SEG 02.09.07.008.00.00 "Genomic Selection in Dairy Cattle in Brazil," CNPq PVE 407246/2013-4 "Genomic Selection in Dairy Gyr and Girolando Breeds," and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) CVZ PPM 00395/14 "Genomic Selection in Brazilian Dairy Breeds" appropriated projects. JBC was supported by appropriated project 1265-31000-096-00, "Improving Genetic Predictions in Dairy Animals Using Phenotypic and Genomic Information," of the Agricultural Research Service of the USDA. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

REFERENCES

Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Gen. Res. https://doi.org/10.1101/gr.094052.109.

Berry, D. P., M. C. Mcclure, and M. P. Mullen. 2014. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. J. Anim. Breed. Genet. 131:165–172. https://doi.org/10.1111/jbg.12067.

Boison, S. a., H. H. R. Neves, M. Pérez O'Brien, Y. T. Utsunomiya, R. Carvalheiro, M. V. G. B. da Silva, J. Sölkner, and J. F. Garcia. 2014. Imputation of non-genotyped individuals using genotyped progeny in Nellore, a *Bos indicus* cattle breed. Livest. Sci. 166:176–189. https://doi.org/10.1016/j.livsci.2014.05.033.

Boison, S. A., D. J. A. Santos, A. H. T. Utsunomiya, R. Carvalheiro, H. H. R. Neves, A. M. P. O'Brien, J. F. Garcia, J. Sölkner, and M. V. G. B. da Silva. 2015. Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (*Bos indicus*) dairy cattle: Comparison of commercially available SNP chips. J. Dairy Sci. 98:4969–4989. https://doi.org/10.3168/jds.2014-9213.

Bolormaa, S., K. Gore, J. H. J. Van Der Werf, B. J. Hayes, and H. D. Daetwyler. 2015. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Anim. Genet. 46:544–556. https://doi.org/10.1111/age.12340.

Carvalheiro, R., S. A. Boison, H. H. R. Neves, M. Sargolzaei, F. S. Schenkel, Y. T. Utsunomiya, A. M. P. O'Brien, J. Sölkner, J. C. McEwan, C. P. Van Tassell, T. S. Sonstegard, and J. F. Garcia. 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46:69. https://doi.org/10.1186/s12711-014-0069-1.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4:7. https://doi.org/10.1186/s13742-015-0047-8.

Chud, T. C. S., R. V. Ventura, F. S. Schenkel, R. Carvalheiro, M. E. Buzanskas, J. O. Rosa, M. de A. Mudadu, M. V. G. B. da Silva, F. B. Mokry, C. R. Marcondes, L. C. A. Regitano, and D. P. Munari. 2015. Strategies for genotype imputation in composite beef cattle. BMC Genet. 16:99. https://doi.org/10.1186/s12863-015-0251-7.

Cleveland, M. A., and J. M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci. 91:3583–3592. https://doi.org/10.2527/jas.2013-6270.

Cole, J. B., and M. V. G. B. da Silva. 2016. Invited Review: Genomic selection in multi-breed dairy cattle populations. Rev. Bras. Zootec. 45:195–202. https://doi.org/10.1590/S1806-92902016000400008.

Cole, J. B., M. Vinicius, and G. Barbosa. 2016. Invited review: Genomic selection in multi-breed dairy cattle populations. Rev. Bras. Zootec. 45:195–202. https://doi.org/10.1590/S1806-92902016000400008.

Druet, T., C. Schrooten, and A. P. W. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93:5443–5454. https://doi.org/10.3168/jds.2010-3255.

García-Ruiz, A., J. Cole, P. Vanraden, G. Wiggans, F. Ruiz, and C. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. USA 113:E3995–E4004.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257. https://doi.org/10.1007/s10709-008-9308-0.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443. https://doi.org/10.3168/jds.2008-1646.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654–663. https://doi.org/10.2135/cropsci2011.07.0358.

Howie, B., J. Marchini, and M. Stephens. 2011. Genotype imputation with thousands of genomes. G3 (Bethesda) 1:457–470. https://doi.org/10.1534/g3.111.001198.

Jattawa, D., M. A. Elzo, S. Koonawootrittriron, and T. Suwanasopee. 2016. Imputation accuracy from low to moderate density single nucleotide polymorphism chips in a Thai multibreed dairy cattle population. Asian-australas. J. Anim. Sci. 29:464–470. https://doi.org/10.5713/ajas.15.0291.

Judge, M. M., J. F. Kearney, M. C. McClure, R. D. Sleator, and D. P. Berry. 2016. Evaluation of developed low-density genotype panels for imputation to higher density in independent dairy and beef cattle populations. J. Anim. Sci. 94:949–962. https://doi.org/10.2527/jas.2015-0044.

Khatkar, M. S., G. Moser, B. J. Hayes, and H. W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13:538. https://doi.org/10.1186/1471-2164-13-538.

Larmer, S. G., M. Sargolzaei, and F. S. Schenkel. 2014. Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. J. Dairy Sci. 97:3128–3141. https://doi.org/10.3168/jds.2013-6826.

Lund, M.S., G. Su, L. Janss, B. Guldbrandtsen, and R. F. BrØndum. 2014. Invited review: Genomic evaluation of cattle in a multi-breed context. Livest. Sci. 166:101–110. https://doi.org/10.1016/j.livsci.2014.05.008.

Moghaddar, N., K. P. Gore, H. D. Daetwyler, B. J. Hayes, and J. H. J. van der Werf. 2015. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. Genet. Sel. Evol. 47:97. https://doi.org/10.1186/s12711-015-0175-8.

Olson, K. M., P. M. VanRaden, and M. E. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 95:5378–5383. https://doi.org/10.3168/jds.2011-5006.

Pausch, H., B. Aigner, R. Emmerling, C. Edel, K.-U. Götz, and R. Fries. 2013. Imputation of high-density genotypes in the Fleckvieh

cattle population. Genet. Sel. Evol. 45:3. https://doi.org/10.1186/1297-9686-45-3.

Piccoli, M. L., J. Braccini, F. F. Cardoso, M. Sargolzaei, S. G. Larmer, and F. S. Schenkel. 2014. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. BMC Genet. 15:157. https://doi.org/10.1186/s12863-014-0157-9.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. https://doi.org/10.1186/1471-2164-15-478.

Thomasen, J. R., A. C. Sørensen, G. Su, P. Madsen, M. S. Lund, and B. Guldbrandtsen. 2013. The admixed population structure in Danish Jersey dairy cattle challenges accurate genomic predictions. J. Anim. Sci. 91:3105–3112. https://doi.org/10.2527/jas.2012-5490.

van Binsbergen, R., M. C. Bink, M. P. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 46:41. https://doi.org/10.1186/1297-9686-46-41.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43:10. https://doi.org/10.1186/1297-9686-43-10.

Ventura, R. V., D. Lu, F. S. Schenkel, Z. Wang, C. Li, and S. P. Miller. 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J. Anim. Sci. 92:1433–1444. https://doi.org/10.2527/jas.2013-6638.

Ventura, R. V., S. P. Miller, K. G. Dodds, B. Auvray, M. Lee, M. Bixley, S. M. Clarke, and J. C. McEwan. 2016. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. Genet. Sel. Evol. 48:71. https://doi.org/10.1186/s12711-016-0244-7.

Weng, Z., Z. Zhang, Q. Zhang, W. Fu, S. He, and X. Ding. 2013. Comparison of different imputation methods from low- to high-density panels using Chinese Holstein cattle. Animal 7:729–735. https://doi.org/10.1017/S1751731112002224.

Wiggans, G. R., P. M. Vanraden, and T. A. Cooper. 2011. The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. 94:3202–3211. https://doi.org/10.3168/jds.2010-3866.

Wray, N. R. 2005. Allele frequencies and the r^2 measure of linkage disequilibrium: Impact on design and interpretation of association studies. Twin Res. Hum. Genet. 8:87–94. https://doi.org/10.1375/1832427053738827.

Xiang, T., P. Ma, T. Ostersen, A. Legarra, and O. F. Christensen. 2015. Imputation of genotypes in Danish purebred and two-way crossbred pigs using low-density panels. Genet. Sel. Evol. 47:54. https://doi.org/10.1186/s12711-015-0134-4.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol. 10:R42. https://doi.org/10.1186/gb-2009-10-4-r42.