J. Dairy Sci. 107:398–411
https://doi.org/10.3168/jds.2023-23250

Imputation accuracy from low- to medium-density SNP chips for US crossbred dairy cattle

Vanille Déru,[1]* Francesco Tiezzi,[2] Paul M. VanRaden,[3] Emmanuel A. Lozada-Soto,[1] Sajjad Toghiani,[3] and Christian Maltecca[1]
[1]Department of Animal Science, North Carolina State University, Raleigh, NC 27607
[2]Department of Agriculture, Food, Environment and Forestry, University of Florence, Florence, 50144, Italy
[3]USDA, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, MD 20705-2350

ABSTRACT

This study aimed at evaluating the quality of imputation accuracy (IA) by marker (IA_m) and by individual (IA_i) in US crossbred dairy cattle. Holstein × Jersey crossbreds were used to evaluate IA from a low- (7K) to a medium-density (50K) SNP chip. Crossbred animals, as well as their sires (53), dams (77), and maternal grandsires (63), were all genotyped with a 78K SNP chip. Seven different scenarios of reference populations were tested, in which some scenarios used different family relationships and others added random unrelated purebred and crossbred individuals to those different family relationship scenarios. The same scenarios were tested on Holstein and Jersey purebred animals to compare these outcomes against those attained in crossbred animals. The genotype imputation was performed with findhap (version 4) software (VanRaden, 2015). There were no significant differences in IA results depending on whether the sire of imputed individuals was Holstein and the dam was Jersey, or vice versa. The IA increased significantly with the addition of related individuals in the reference population, from 86.70 ± 0.06% when only sires or dams were included in the reference population to 90.09 ± 0.06% when sire (S), dam (D), and maternal grandsire genomic data were combined in the reference population. In all scenarios including related individuals in the reference population, IA_m and IA_i were significantly superior in purebred Jersey and Holstein animals than in crossbreds, ranging from 90.75 ± 0.06 to 94.02 ± 0.06%, and from 90.88 ± 0.11 to 94.04 ± 0.10%, respectively. Additionally, a scenario called S_{PB}+D_{LD} (where PB indicates purebread and LD indicates low density), similar to the genomic evaluations performed on US crossbred dairy, was tested. In this scenario, the information from the 5 evaluated breeds (Ayrshire, Brown Swiss, Guernsey, Holstein, and Jersey) genotyped with a 50K SNP chip and genomic information from the dams genotyped with a 7K SNP chip were combined in the reference population, and the IA_m and IA_i were 80.87 ± 0.06% and 80.85 ± 0.08%, respectively. Adding randomly nonrelated genotyped individuals in the reference population reduced IA for both purebred and crossbred cows, except for scenario S_{PB}+D_{LD}, where adding crossbreds to the reference population increased IA values. Our findings demonstrate that IA for US Holstein × Jersey crossbred ranged from 85 to 90%, and emphasize the significance of designing and defining the reference population for improved IA.

Key words: dairy crossbred, imputation accuracy, SNP chip

INTRODUCTION

Low-density SNP chips are employed in dairy populations to genotype the female population. In contrast, medium-density chips (50K SNPs or more), which are expensive, are typically reserved for valuable individuals, mostly contributing to the male path of selection (Wiggans et al., 2012). For these low-density genotyped individuals, imputation to medium density is performed before including information in genomic prediction models (VanRaden et al., 2013a). This genotyping strategy has been highly successful, and now the dairy population in the Council on Dairy Cattle Breeding (**CDCB**) database contains more than 6 million genotyped individuals. High imputation accuracy is crucial for estimating the most accurate genetic values of animals, as reviewed in Calus et al. (2014). In purebred dairy cattle, several studies have shown that imputation accuracy (**IA**) from low- to medium- or high-density SNP chips is high (Weigel et al., 2010b; Mulder et al., 2012; VanRaden et al., 2013b), and accuracy of genomic selection is not notably affected by imputation (Weigel et al., 2010a; Wiggans et al., 2012; VanRaden et al., 2013b). However, most of these studies have been conducted in purebred dairy cattle.

Received January 10, 2023.
Accepted June 16, 2023.
*Corresponding author: vanillederu@hotmail.fr

In addition to heterosis and breed complementarity, crossbreeding is a potential approach to improve cow fertility, cow health, calf survival, and disease resistance (Heins et al., 2008; Sørensen et al., 2008). Crossbreeding in dairy cattle has increased in popularity in recent years. In the United States, almost 7% of the 3.7 million milk-recorded cattle were crossbred in 2020 (Norman et al., 2021; Wiggans et al., 2021). Currently, the number of crossbred dairy cattle with genomic information in the CDCB database is increasing, with more than 50K individuals, and it continues to grow (Figure 1). Genomic data of crossbred cows contribute to the genomic prediction of sires, and the genomic evaluation of crossbred cattle in the United States is currently performed with inclusion of the 5 evaluated breeds (Ayrshire, Brown Swiss, Guernsey, Holstein, and Jersey) genotyped with medium-density SNP chips and dams genotyped with low-density SNP chips (Wiggans et al., 2021). However, it is not known how IA of crossbred dairy cattle is different from their purebred counterparts. To our knowledge, only 2 studies have estimated IA for dairy crossbred populations, one in African dairy cattle population (Aliloo et al., 2018) and the other in a Girolando (Gyr × Holstein) population (Oliveira Júnior et al., 2017). To date, no study has yet reported results of IA in US crossbred cows.

The aims of this study were as follows: to assess IA in US crossbred dairy cattle from a low-density to a medium-density SNP chip, using different defined scenarios to evaluate the effects of reference population size and its relationship with imputed crossbred dairy population, to compare IA values obtained in crossbred and purebred dairy populations, to compare the proportions of individuals and markers based on their imputation accuracy among scenarios, and to compare IA for each chromosome under different defined scenarios.

MATERIALS AND METHODS

No live animals were used in this study, and therefore Institutional Animal Care and Use Committee approval was not required. This study focused on crossbreds Holstein × Jersey (**HO×JE**, Holstein sire and Jersey dam) and Jersey × Holstein (**JE×HO**, Jersey sire and Holstein dam) dairy populations, since these types of crossing are the most popular ones in the US. In the CDCB database, 79% of the genotyped US crossbred dairy population were HO×JE or JE×HO (Table 1).

Animals and Data

To investigate IA under different genomic information availability, 110 F_1 crossbreds HO×JE and JE×HO genotyped with a 78K GGP-HD chip (GeneSeek Genomic Profiler, >77K markers; Neogen Corporation; **78K**) were chosen based on available pedigree and genomic information. The breed base representations (**BBR**) of crossbred were 50 ± 10% Holstein and 50 ± 10% Jersey. The BBR estimated the similarity of the alleles present in the 5 purebred reference groups against those of individuals genotyped. These values were restricted to be between 0 and 100% for each genotyped animal and summed, as explained in VanRaden et al. (2020) and Wiggans et al. (2021). These animals' sires, dams, and maternal grandsires were genotyped with the same SNP chip (Supplemental Table S1, https://doi.org/10.6084/m9.figshare.21858864; Déru, 2023). Of these 110 crossbreds, 53 were HO×JE and 57 were JE×HO.

Genotypes

Before imputation, real genotypes were phased within each purebred and crossbred population for all animals with Beagle 5.3 (Browning et al., 2021). Then, markers of the 78K SNP chip were masked to mimic the Illumina Bovine LD v2.0 BeadChip chip (7,931 markers, Illumina Inc.; **7K SNP chip**) in the validation population. Similarly, markers were masked to mimic the BovineSNP50 Genotyping BeadChip chip (54,609 markers, Illumina Inc.; **50K SNP chip**) in the reference population. In total, 6,912 SNP chip overlapped with the 50K SNP chip. These 2 SNP chips were studied because they are the most commonly used low- and medium-density chips for predicting the genetic merit of dairy cattle (Schefers and Weigel, 2012).

In previous studies conducted in crossbred beef and dairy cows, the genetic connectedness between the reference and the validation population has shown an influence on IA values (Ventura et al., 2014; Chud et al., 2015; Oliveira Júnior et al., 2017; Aliloo et al., 2018). Therefore, a measure of genomic relationship between animals in reference and validation sets was calculated within each scenario to investigate the effect of connectedness between reference and imputed populations on IA values. For this purpose, the genomic relationship matrix (**G**) was calculated following the first method of VanRaden (VanRaden, 2008). Then, the averaged genomic value (extracted from the **G** matrix) between all individuals included in the reference and validation population were estimated within each scenario and called the genomic relationship coefficient (**GRC**).

Imputation Scenarios

Different reference population scenarios for imputing US crossbred cattle were created and are summarized in Table 2. Seven scenarios differed based on the relatives' information between reference and validation

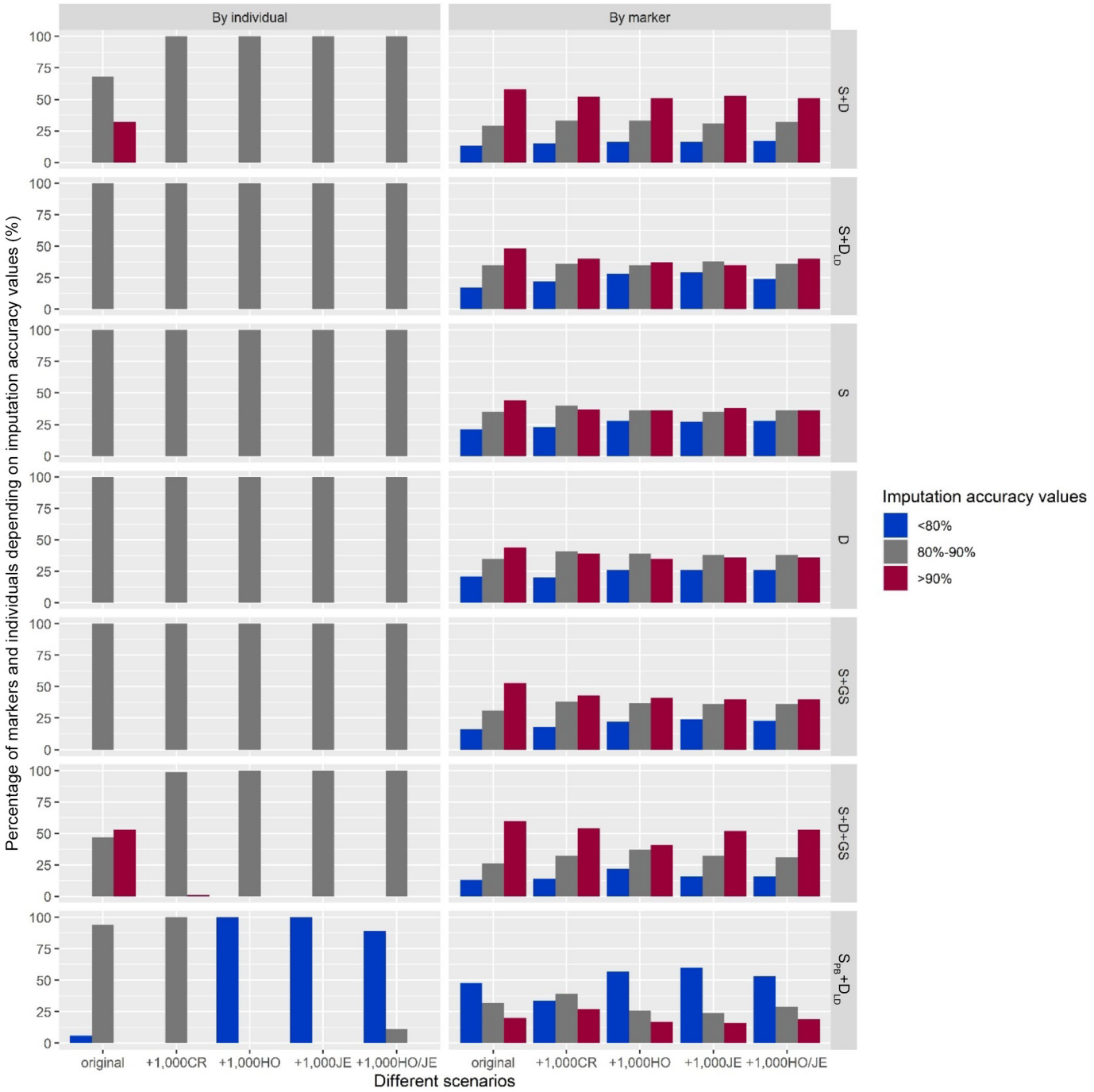

Figure 1. Percentage of individuals and markers according to the imputation accuracy values in the 8 different scenarios presented in Table 2 (original) and with the addition, in each scenario, of 1,000 crossbred animals (+1,000CR), 1,000 purebred Holstein dairy cattle (+1,000HO) and 1,000 Jersey dairy cattle (+1,000JE) separately and combined (+1,000HO/JE). S+D = sires and dams genotyped with a 50K chip in reference population; $S+D_{LD}$ = sires genotyped with a 50K chip and dams genotyped with a 7K chip in reference population; D = dams genotyped with a 50K chip in reference population; S+GS = sires and maternal grandsires genotyped with a 50K chip in reference population; S+D+GS = sires, dams, and maternal grandsires genotyped with a 50K chip in reference population; $S_{PB}+D_{LD}$ = purebred bulls from 5 different bulls genotyped with a 50K chip and dams genotyped with a 7K chip in the reference population.

Table 1. Number of genotyped US crossbred dairy cattle (in bold) and those with available genomic data for the 78K marker chip in particular (in italics) depending on the sire breed in row and dam breed in column in the Council of Dairy Cattle Breeding database

	Dam breed					
Sire breed	Holstein	Jersey	Ayrshire	Brown Swiss	Guernsey	Crossbred
Holstein	—	**19,641**	**64**	**102**	**15**	**3,750**
		94	*2*	*1*	*0*	*92*
Jersey	**21,541**	—	**23**	**14**	**7**	**5,897**
	132		*0*	*0*	*0*	*471*
Ayrshire	**422**	**13**	—	**1**	**0**	**7**
	15	*0*		*0*	*0*	*1*
Brown Swiss	**270**	**46**	**0**	—	**1**	**28**
	2	*1*	*0*		*0*	*0*
Guernsey	**57**	**15**	**0**	**4**	—	**5**
	0	*0*	*0*	*0*		*0*
Crossbred	**45**	**66**	**0**	**0**	**0**	—
	4	*0*	*0*	*0*	*0*	

populations. The population of imputed individuals always consisted of the 110 from each crossbred type (either HO×JE or JE×HO) described above; however, the reference panel changed from one scenario to another. The genotypes indicated in each scenario were those represented in the reference population. The 7 scenarios were the following:

- **S+D**: This scenario included genomic information of the sires (**S**) and the dams (**D**) of crossbred animals genotyped with the 50K SNP chip.
- **S+D$_{LD}$**: This scenario included genomic information of the sires genotyped with the 50K SNP chip and the dams of crossbred animals genotyped with the 7K SNP chip (low density, **LD**).
- **S**: This scenario included genomic information of the sires of crossbred animals genotyped with the 50K SNP chip.
- **D**: This scenario included genomic information of the dams of crossbred animals genotyped with the 50K SNP chip.
- **S+GS**: This scenario included genomic information of the sires and maternal grandsires (**GS**) of crossbred animals genotyped with the 50K SNP chip.
- **S+D+GS**: This scenario included genomic information of the sires, dams, and maternal grandsires of crossbred animals all genotyped with the 50K SNP chip.
- **S$_{PB}$+D$_{LD}$**: This scenario included genomic information of purebred (**PB**) bulls from 5 different breeds (53 Ayrshire, 53 Guernsey, 53 Brown Swiss, 53 Holstein, and 53 Jersey), including sires of crossbred (26 out of the 53 Jersey and 27 out of the 53 Holstein, respectively) genotyped with the 50K SNP chip and the dams of crossbred animals genotyped with the 7K SNP chip.

The objective of the S$_{PB}$+D$_{LD}$ scenario was to be similar to the genomic evaluations currently performed by the CDCB (Wiggans et al., 2021). The purebred bulls were randomly chosen based on whether they were genotyped with a 78K SNP chip, had a BBR of 100%, and were born after 2008. Among the 53 Jersey and 53 Holstein, 26 and 27, respectively, were the sires of the crossbreds. The random selection of these animals was repeated 5 times. The mean of the 5 replicates was subsequently employed as a metric for subsequent IA calculations and shown in this scenario.

In addition, we evaluated the effects of adding purebred or crossbred genomic information in the reference population. For all aforementioned scenarios, the effects on IA of adding to the reference panel 1,000 random purebred Holsteins or Jerseys, or a combination of the 2, as well as 1,000 crossbred animals were evaluated. For purebred cattle, all randomly chosen individuals were genotyped with a 78K SNP chip, and their genomic information was masked to mimic the 50K SNP chip. They were not the parents or grandparents of the crossbreds in the validation population. Because insufficient HO×JE crossbreds were genotyped with the 78K SNP chip, the selection of crossbred animals was expanded to include any crossbred composed of 2 or more of the main US breeds (Guernsey, Ayrshire, Brown Swiss, Jersey, or Holstein). A total of 4,318 crossbred animals met these requirements. The sampling of 1,000 random individuals was repeated 5 times (with replacement). The average of the 5 replicates was subsequently employed as a metric for subsequent calculation, as shown in the results.

Finally, purebred Holstein or Jersey cattle were used in the validation population for the first 6 reference population scenarios, serving as a benchmark for IA (Supplemental Table S2, https://doi.org/10.6084/m9.figshare.21858906.v1; Déru et al., 2023a).

Table 2. Imputation scenarios used in the study for 110 Holstein × Jersey (HO×JE) and Jersey × Holstein (JE×HO) crosses[1]

Scenario (abbreviation)	No. reference	Description of reference	Description of crossbreds
Sire + dam (S+D)	130 (53 sires + 77 dams)	Dams and sires genotyped with a 50K chip	53 HO×JE, 57 JE×HO
Sire + dam_{LD} ($S+D_{LD}$)	130 (53 sires + 77 dams)	Sires genotyped with a 50K chip, dams genotyped with a 7K chip	53 HO×JE, 57 JE×HO
Sire only (S)	53	Sires genotyped with a 50K chip	53 HO×JE, 57 JE×HO
Dam only (D)	77	Dams genotyped with a 50K chip	53 HO×JE, 57 JE×HO
Sire + maternal grandsire (S+GS)	115 (53 sires + 62 maternal grandsires)	Sires and maternal grandsires genotyped with a 50K chip	53 HO×JE, 57 JE×HO
Sire + dam + maternal grandsire (S+D+GS)	192 (77 dams + 62 maternal grandsires + 53 sires)	Sires, dams, and maternal grandsires genotyped with a 50K chip	53 HO×JE, 57 JE×HO
$Sire_{PB}$ + dam_{LD} ($S_{PB}+D_{LD}$)	342 (53 Holstein bulls + 53 Jersey bulls + 53 Guernsey bulls + 53 Ayrshire bulls + 53 Brown Swiss bulls + 77 dams)	Purebred bulls included sires of crossbreds genotyped with a 50K chip and dams genotyped with a 7K chip	53 HO×JE, 57 JE×HO

[1]HO = Holstein; JE = Jersey; S = sire; D = dam; GS = maternal grandsire; PB = purebred; LD = low density. HO×JE: sire is Holstein, dam is Jersey. JE×HO: dam is Jersey, sire is Holstein.

Imputation Accuracy

Genotype imputation was performed using findhap software (version 4; VanRaden, 2015). The latest version of findhap was used; this version is currently not quite ready for routine use on US chip data but performs better than version 3 for sequence data. The parameter settings were those recommended to combine 3K or 6K with 50K SNP chip (VanRaden, 2015). The maximum and minimum lengths of segments to check for haplotyping were 600 and 75 SNPs, respectively. The number of steps to get from maximum to minimum length was 3; the maximum iterations for haplotype imputation was 4; and the maximum number of different haplotypes within any segment was fixed at 1,000. The fraction of miscalled alleles was set to be 0.004.

The IA per marker ($\mathbf{IA_m}$) and per individual ($\mathbf{IA_i}$) were calculated as the averaged proportion of correctly imputed genotypes between actual and imputed genotypes per marker and per individual (i.e., concordance rate, as presented in Calus et al., 2014), with the SNPs on the 29 autosomal chromosomes considered; thus not on sex chromosomes.

Post-Analysis Statistics

A statistical comparison was conducted to assess the differences in IA_m and IA_i between the different scenarios. A linear model was fitted for IA_m:

$$y = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where y is IA_m; $\boldsymbol{\beta}$ is the vector of fixed effects: scenarios and interaction between chromosomes and scenarios; $\mathbf{X}$ is the incidence matrix relating observations to fixed effects; and $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma_e^2\right)$ is the residual random effect, with $\mathbf{I}$ the incidence matrix and σ_e^2 the residual variance.

A similar model for IA_i was fitted, including only the effect of scenarios. Least squares means (**LSM**) for IA_m and IA_i were obtained for each scenario using the *lsmeans()* function of the lsmeans package in R (Lenth, 2016). The LSM were compared between scenarios with a *t*-test via the *contrast()* function of the same package. The distributions of IA_m and IA_i were obtained for all scenarios. Individuals and markers were classified in 3 classes according to their IA values: less than 80% (low), between 80 and 90% (moderate), and above 90% (high). To compare the proportion of individuals or markers between groups and scenarios, a 2-proportions Z-test was performed with the *prop.test()* function in R (R Core Team, 2016).

To observe any differences in IA_m for a given chromosome between scenarios, the difference in IA_m between scenarios for each chromosome were compared. The contrasts between LSM obtained between scenarios were estimated with the *contrast()* function in R (Lenth, 2016), which allows pairwise comparisons of LSM by testing linear contrasts among predictions. To evaluate the effects and the significance of the length of the chromosome on IA values, the Pearson correlation between the length of the chromosome and IA was calculated for each scenario with the *cor.test()* function in R (R Core Team, 2016). Similarly, the

Table 3. Least squares means of imputation accuracy per marker and per individual in the different scenarios for Holstein × Jersey (HO×JE), Jersey × Holstein (JE×HO), both combined (Both), purebred Holstein (HO), and purebred Jersey (JE), along with their standard errors

	LSM[2] (SE), imputation accuracy by marker					LSM[3] (SE), imputation accuracy by individual				
Scenario[1]	HO×JE	JE×HO	Both	HO	JE	HO×JE	JE×HO	Both	HO	JE
S+D	89.70[a]	89.66[a]	89.68[a]	94.02[a]	93.94[a]	89.65[a]	89.59[a]	89.62[a]	94.04[a]	94.01[a]
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.10)	(0.12)	(0.08)	(0.10)	(0.11)
S+D_{LD}	87.69[b]	87.58[a]	87.64[b]	92.80[b]	92.73[b]	87.71[b]	87.53[b]	87.62[b]	92.89[b]	92.82[b]
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.10)	(0.12)	(0.08)	(0.10)	(0.11)
S	86.79[c]	86.61[b]	86.70[c]	91.18[c]	91.01[c]	86.85[c]	86.64[c]	86.74[c]	91.26[c]	91.15[c]
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.10)	(0.12)	(0.08)	(0.10)	(0.11)
D	86.66[c]	86.99[c]	86.83[c]	91.36[c]	91.35[d]	86.51[c]	86.88[c]	86.70[c]	91.53[c]	91.54[c]
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.10)	(0.12)	(0.08)	(0.10)	(0.11)
S+GS	88.42[d]	88.29[d]	88.35[d]	91.28[c]	90.75[e]	88.49[d]	88.39[d]	88.44[d]	91.27[c]	90.88[c]
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.10)	(0.12)	(0.08)	(0.10)	(0.11)
S+D+GS	90.08[e]	90.10[e]	90.09[e]	93.65[d]	92.99[f]	89.99[a]	89.98[a]	89.99[e]	93.74[a]	93.11[b]
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.10)	(0.12)	(0.08)	(0.10)	(0.11)
S_{PB}+D_{LD}	80.91[f]	80.84[f]	80.87[f]	—	—	80.91[e]	80.79[e]	80.85[f]	—	—
	(0.06)	(0.06)	(0.06)				(0.12)	(0.08)		

[a–f]Least squares means in the same column with different superscripts are statistically different according to a *t*-test ($P < 0.05$).

[1]S+D = sires and dams genotyped with a 50K chip in reference population; S+D_{LD} = sires genotyped with a 50K chip and dams genotyped with a 7K chip in reference population; S = sire; D = dams genotyped with a 50K chip in reference population; S+GS = sires and maternal grandsires genotyped with a 50K chip in reference population; S+D+GS = sires, dams, and maternal grandsires genotyped with a 50K chip in reference population; S_{PB}+D_{LD} = purebred bulls from 5 different bulls genotyped with a 50K chip and dams genotyped with a 7K chip in the reference population.

[2]From a linear mixed model including the fixed effects of the scenario, the chromosome, and the interaction between the scenario and the chromosome. All effects were significant for $P < 0.0001$.

[3]From a linear mixed model including the fixed effects of the scenario, which was significant for $P < 0.0001$. The effect of the type of crossbred (HO×JE and JE×HO) was tested but not included in the model because this was not significant ($P = 0.65$).

Pearson correlation between the percentage of homozygosity per marker and the IA_m obtained was also calculated.

RESULTS

Comparison of Imputation Accuracy by Marker and Individual

The LSM of IA_m and IA_i in the different scenarios for HO×JE and JE×HO separately, combined, and within purebred animals, are presented in Table 3.

In all scenarios, we found no significant difference between the imputation accuracy (IA_m and IA_i) between HO×JE, JE×HO, and both type of crossbred combined ($P = 0.65$). Thus, only results for the combined population are shown and those found in purebreds. In scenarios S and D, IA was 86.70 ± 0.06 and 86.83 ± 0.06 for crossbred but significantly higher in purebred animals ($P < 0.05$; >91.01 ± 0.06). No significant differences were observed in crossbreds for IA_m and IA_i, regardless of whether the genomic information of the sire or the dam was included in the reference population. In the S+D scenario, IA was significantly higher than in scenarios S and D ($P < 0.05$), with values that were still higher in purebreds than in crossbreds (around +4 percentage points higher in purebreds). In the scenario S+D_{LD}, IA values were between those of scenario S+D and scenarios S and D. In scenario S+GS, IA_m and IA_i were lower than in the S+D scenario for crossbreds (−1.37 percentage points for IA_m and −1.18 percentage points for IA_i; $P < 0.05$). The same observation was made in purebred populations, but IA values were still significantly higher than in crossbred populations for this scenario ($P < 0.05$). In the scenarios including related individuals in the reference population, the lowest IA was observed for the scenario S_{PB}+D_{LD} ($IA_m = 80.87 \pm 0.06$ and $IA_i = 80.85 \pm 0.08$). Finally, the highest IA_m and IA_i were found for the scenario that combined the genomic information of sires, dams, and maternal grandsires in the reference (scenario S+D+GS) in all crossbreds ($IA_m = 90.09 \pm 0.06$ and $IA_i = 89.99 \pm 0.08$), purebred Jerseys ($IA_m = 92.99 \pm 0.10$ and $IA_i = 93.11 \pm 0.11$), and purebred Holsteins ($IA_m = 93.65 \pm 0.06$ and $IA_i = 93.74 \pm 0.06$).

In all scenarios with the addition of related individuals (S, D, S+D, S+D_{LD}, S+GS, and S+D+GS), IA values were significantly higher in purebred animals than in

Table 4. Percentage increase in imputation accuracy by marker and by individual in different scenarios with 110 crossbreds (CR), purebred Holstein (HO), or purebred Jersey (JE), and with the addition of 1,000 random purebred Holstein (+1,000 HO), Jersey (+1,000 JE), crossbred (+1,000 CR), or +1,000 Holstein and +1,000 Jersey combined (+1,000 HO/JE) in the reference population

	Percentage increase of imputation accuracy per marker[2]				Percentage increase of imputation accuracy per individual[2]			
Scenario[1]	+1,000 HO	+1,000 JE	+1,000 CR	+1,000 HO/JE	+1,000 HO	+1,000 JE	+1,000 CR	+1,000 HO/JE
Crossbred								
S+D	−2%*	−1%*	−2%*	−2%*	−2%*	−2%*	−2%*,	−2%*
S+D_{LD}	−3%*	−3%*	−2%*	−2%*	−3%*	−3%*	−2%*	−2%*
S	−2%*	−1%*	−1%*	−2%*	2%*	−2%*	−1%*	−2%*
D	−1%*	−2%*	−1%*	−2%*	−1%*	−2%*	−1%*	−2%*
S+GS	−2%*	−3%*	−4%*	−3%*	−3%*	−3%*	−4%*	−3%*
S+D+GS	−2%*	−2%*	−2%*	−2%*	−2%*	−2%*	−2%*	−2%*
S_{PB}+D_{LD}	−2%*	−3%*	+3%*	−1%*	−2%*	−3%*	+3%*	−2%*
Purebred Holstein								
S+D	−6%*	—	—	—	−6%*	—	—	—
S+D_{LD}	6%*	—	—	—	−8%**	—	—	—
S	−4%*	—	—	—	−4%*	—	—	—
D	−4%*	—	—	—	−4%*	—	—	—
S+GS	−4%*	—	—	—	−4%*	—	—	—
S+D+GS	−6%*	—	—	—	−6%*	—	—	—
Purebred Jersey								
S+D	—	−3%*	—	—	—	−3%*	—	—
S+D_{LD}	—	−2%*	—	—	—	−2%*	—	—
S	—	−1%*	—	—	—	−1%*	—	—
D	—	−1%*	—	—	—	−1%*	—	—
S+GS	—	−2%**	—	—	—	−0%	—	—
S+D+GS	—	−2%*	—	—	—	−2%*	—	—

[1]S+D = sires and dams genotyped with a 50K chip in reference population; S+D_{LD} = sires genotyped with a 50K chip and dams genotyped with a 7K chip in reference population; D = dams genotyped with a 50K chip in reference population; S+GS = sires and maternal grandsires genotyped with a 50K chip in reference population; S+D+GS = sires, dams, and maternal grandsires genotyped with a 50K chip in reference population; S_{PB}+D_{LD} = purebred bulls from 5 different bulls genotyped with a 50K chip and dams genotyped with a 7K chip in the reference population.

[2]Least squares of imputation accuracy were compared before and after the addition of supplementary dairy cattle in the reference population by a *t*-test.

*$P < 0.05$. **$P < 0.01$.

crossbred animals, and adding more related individuals to the reference population increased IA values.

Effects of Extended Reference Population With Unrelated Individuals on Imputation Accuracy

Table 4 highlights the effects of adding 1,000 unrelated individuals of crossbreds and purebreds to the reference population on IA_m and IA_i in the corresponding validation crossbred or purebred populations.

For scenarios S, D, S+D, S+D_{LD}, S+GS, and S+D+GS, we detected no increase of IA_m and IA_i with adding 1,000 unrelated individuals to crossbreds and purebred in the reference population. A decrease of IA in all scenarios (from −1% to −6%) was observed with the addition of these individuals. However, for S_{PB}+D_{LD}, the inclusion of 1,000 crossbreds to the reference population increased IA_m (+3%) and IA_i (+3%), but the inclusion of purebreds did not increase IA values.

In original scenarios (S, D, S+D, S+D_{LD}, S+GS, and S+D+GS) without the addition of random individuals, the GRC was between 0.316 and 0.323. With the addition of random purebreds and crossbreds in the reference population, the GRC was lower than in original scenarios (0.287 to 0.305), except in the scenario S_{PB}+D_{LD}, where GRC was lower in original scenario (0.262 to 0.264) than with the addition of supplementary individuals (0.270 to 0.284; Supplemental Table S3, https://figshare.com/articles/figure/Supplemental_Table_S3/21858924; Déru et al., 2023b).

Figures 2a and 2b report the values and percentage of increase of IA_m and IA_i when adding related individuals to the 1,000 unrelated individuals in comparison to a scenario with only these random individuals and unrelated individuals in the reference population.

When no related individuals were included in the reference population, IA were higher when +1,000 random crossbreds were added to the reference population (IA_m = 80.75 and IA_i = 79.14) than when +1,000 HO (IA_m = 75.92 and IA_i = 75.01), +1,000 JE (IA_m = 77.30 and IA_i = 77.77), and 1,000 HO/JE (IA_m = 77.99 and IA_i = 79.14) were added to the reference population

When random individuals were added to the reference population, IA were higher when related individuals were already present in the reference population (scenarios S, D, S+D, S+D_{LD}, S+GS, and S+D+GS)

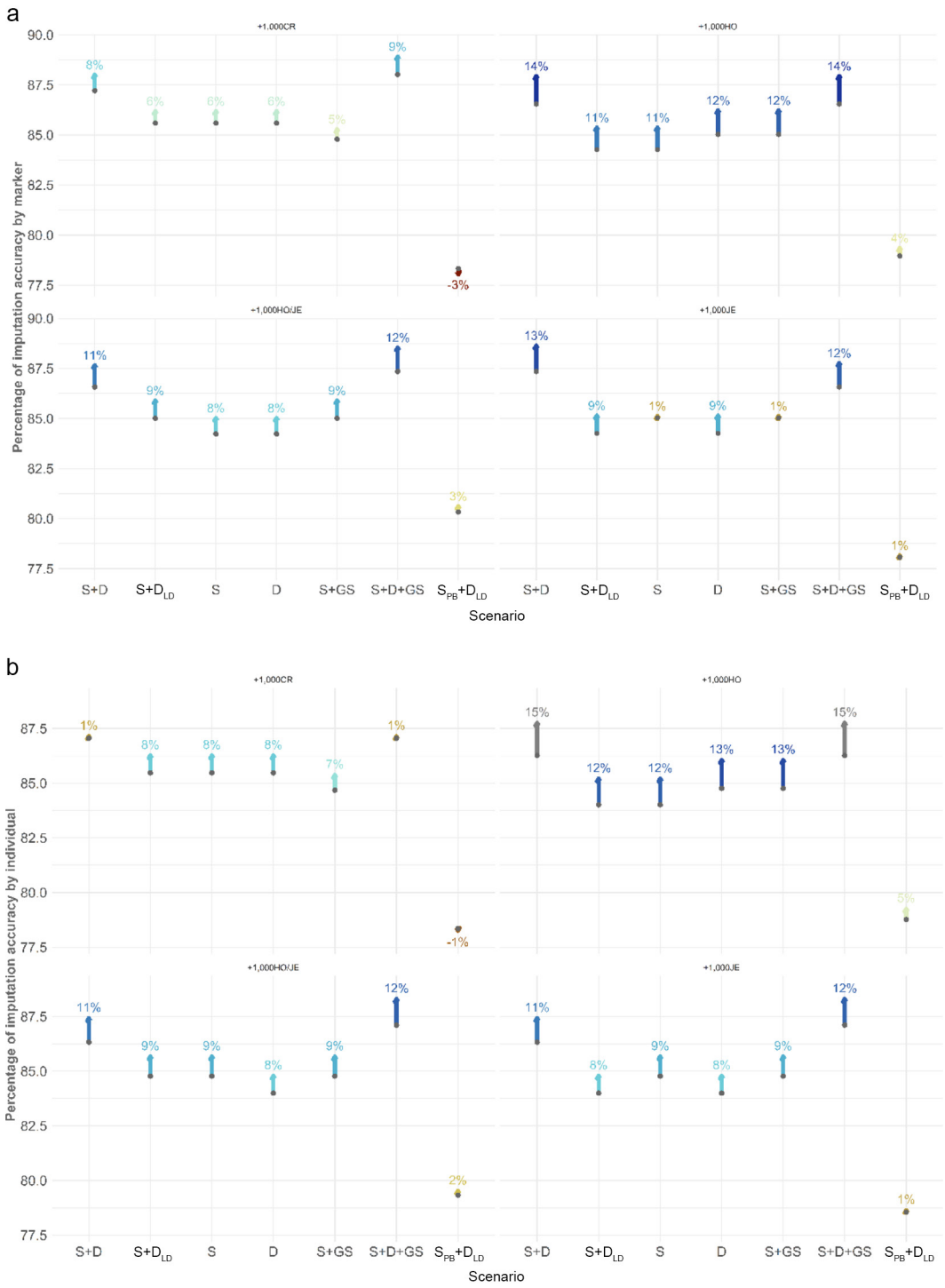

Figure 2. (a) Percentage increase in imputation accuracy by marker in different scenarios compared with the scenario with no related individuals in the reference population, within scenarios with the addition of 1,000 random purebred Holstein (+1,000HO), Jersey (+1,000JE), crossbred (+1,000CR) or +1,000 Holstein and +1,000 Jersey combined (+1,000HO/JE) in the reference population. (b) Percentage increase in imputation accuracy by individual in different scenarios compared with the scenario with no related individuals in the reference population, within scenarios with +1,000HO, +1,000JE, +1,000CR, or +1,000HO/JE in the reference population. S+D = sires and dams genotyped with a 50K chip in the reference population; S+D_{LD} = sires genotyped with a 50K chip and dams genotyped with a 7K chip in reference population; D = dams genotyped with a 50K chip in reference population; S+GS = sires and maternal grandsires genotyped with a 50K chip in reference population; S+D+GS = sires, dams, and maternal grandsires genotyped with a 50K chip in reference population; S_{PB}+D_{LD} = purebred bulls from 5 different bulls genotyped with a 50K chip and dams genotyped with a 7K chip in the reference population. Blue-red color represents positive values in blue, negative values in red, and values close to zero in yellow.

compared with when no related individuals were present, in the order of +5 to +15%. However, IA were higher in the scenario with no relatives but 1,000 crossbreds, compared with the $S_{PB}+D_{LD}$ scenario with 1,000 additional crossed individuals (+1 to +3%).

Distribution of Individuals and Markers According to Imputation Accuracy

Figure 1 illustrates the percentages of individuals and markers based on estimated IA values within different imputation scenarios; detailed values are shown in Supplemental Tables S4 (https://doi.org/10.6084/m9.figshare.21858930; Déru et al., 2023c) and S5 (https://doi.org/10.6084/m9.figshare.21858948; Déru et al., 2023d).

The results between crossbred and purebred animals within the imputation scenarios compared and the percentages of individuals and markers on estimated IA values were categorized into 3 groups (IA <80%, 80% < IA < 90%, and IA >90%). Across all original imputation scenarios having relative individuals in the reference population, the percentage of higher imputed markers (IA_m >90%) was significantly higher for purebreds (66–79%) than for crossbreds (44–60%). Additionally, purebred animals had a significantly higher percentage of accurate imputed individuals (>92%) than crossbred animals (0–53%). In all these imputation scenarios, the percentage of lower (IA_m <80%) or medium (80% < IA_m < 90%) imputed markers was higher and observed more in crossbred than in purebred animals.

Furthermore, the results of imputed crossbreds within defined imputation scenarios were compared. As more related individuals were added to the reference population, the percentage of accurately imputed markers and individuals increased. In fact, the percentages of imputed markers greater than 90% in the S, D, $S+D_{LD}$, S+GS, S+D, and S+D+GS scenarios were 44, 44, 48, 53, 58, and 60%, respectively. Similarly, the percentages of individuals well imputed over 90% when imputation scenarios defined for S, D, $S+D_{LD}$, S+GS, S+D, and S+D+GS in the reference populations were 0, 0, 0, 0, 32, and 53%, respectively. However, the percentages of markers and individuals well imputed over 90% in the $S_{PB}+D_{LD}$ scenario were low, with 20% and 0%, respectively.

The addition of 1,000 unrelated genotyped crossbreds to the reference population of scenarios (S, D, $S+D_{LD}$, S+GS, S+D, and S+D+GS) did not improve the percentage of imputed markers and individuals, and in some scenarios this percentage was decreased. A similar pattern was observed when 1,000 unrelated genotyped purebreds (Holstein or Jersey) were added to those predetermined reference population scenarios. In scenario $S_{PB}+D_{LD}$, an increase in IA values was already observed with adding 1,000 unrelated crossbred animals to the reference population. In this case, the percentage of well imputed markers over 90% increased from 20 to 27%, and the percentage of lower imputed markers (IA_m < 80%) decreased from 6 to 0%.

Imputation Accuracy by Chromosome

Variability in IA by chromosome was observed across different imputation scenarios and was similar to overall IA_m across all scenarios. The IA by chromosome ranging between 86.01 and 93.20%, 84.47 to 90.18%, 82.78 to 90.18%, 84.04 to 90.17%, 84.42 to 92.10%, 86.73 to 93.38%, and 78.24 to 82.78% for S+D, $S+D_{LD}$, S, D, S+GS, S+D+GS, and $S_{PB}+D_{LD}$ scenarios, respectively. Lower IA was found for chromosomes 27, 27, 27, 28, 29, 25, and 28, and higher IA was found for chromosomes 26, 26, 6, 19, 20, 19 and 26 in scenarios S+D, $S+D_{LD}$, S, D, S+GS, S+D+GS, and $S_{PB}+D_{LD}$, respectively.

Only 5 imputation scenarios (S+D: 0.38l $P < 0.05$), S: 0.52 ($P < 0.01$), D: 0.50 ($P < 0.01$), S+D+GS: 0.52 ($P < 0.01$), $S_{PB}+D_{LD}$: 0.46 ($P < 0.05$)] revealed a statistically significant Pearson correlation between chromosome length and IA_m. On the contrary, we found a meaningful correlation between marker homozygosity and IA_m, with Pearson correlations greater than 0.82 in all defined imputation scenarios ($P < 0.0001$).

Differences in Chromosome Imputation Accuracy Between Scenarios

Figure 3 depicts the comparison of LSMeans of IA by chromosome between paired imputation scenarios for imputed crossbred animals. Even though the averaged IA_m was not significantly different between the S and D scenarios, the estimated IA by chromosome were significantly different for 13 out of the 29 chromosomes between these 2 scenarios (Chromosomes 2, 4, 6, 10, 12, 15, 18, 19, 22, 23, 24, 27, 28; $P < 0.05$).

With the exception of chromosomes 6, 15, and 22, all chromosomes showed an increase in IA from extending reference population scenario S to S+D. However, expanding reference population scenario D to S+D increased IA on all chromosomes except for chromosomes 23, 27, and 29. Additionally, it was discovered that when the genotypes of the dams were added to the reference population of imputation scenario S+GS, IA was not significantly increased on chromosome 6 but was significantly increased on chromosomes 15 and 22 ($P > 0.05$ and $P > 0.001$, respectively), which was not the case when adding the genotypes of the dams to the reference population scenario S. When maternal grandsire genotypes were added to the reference population

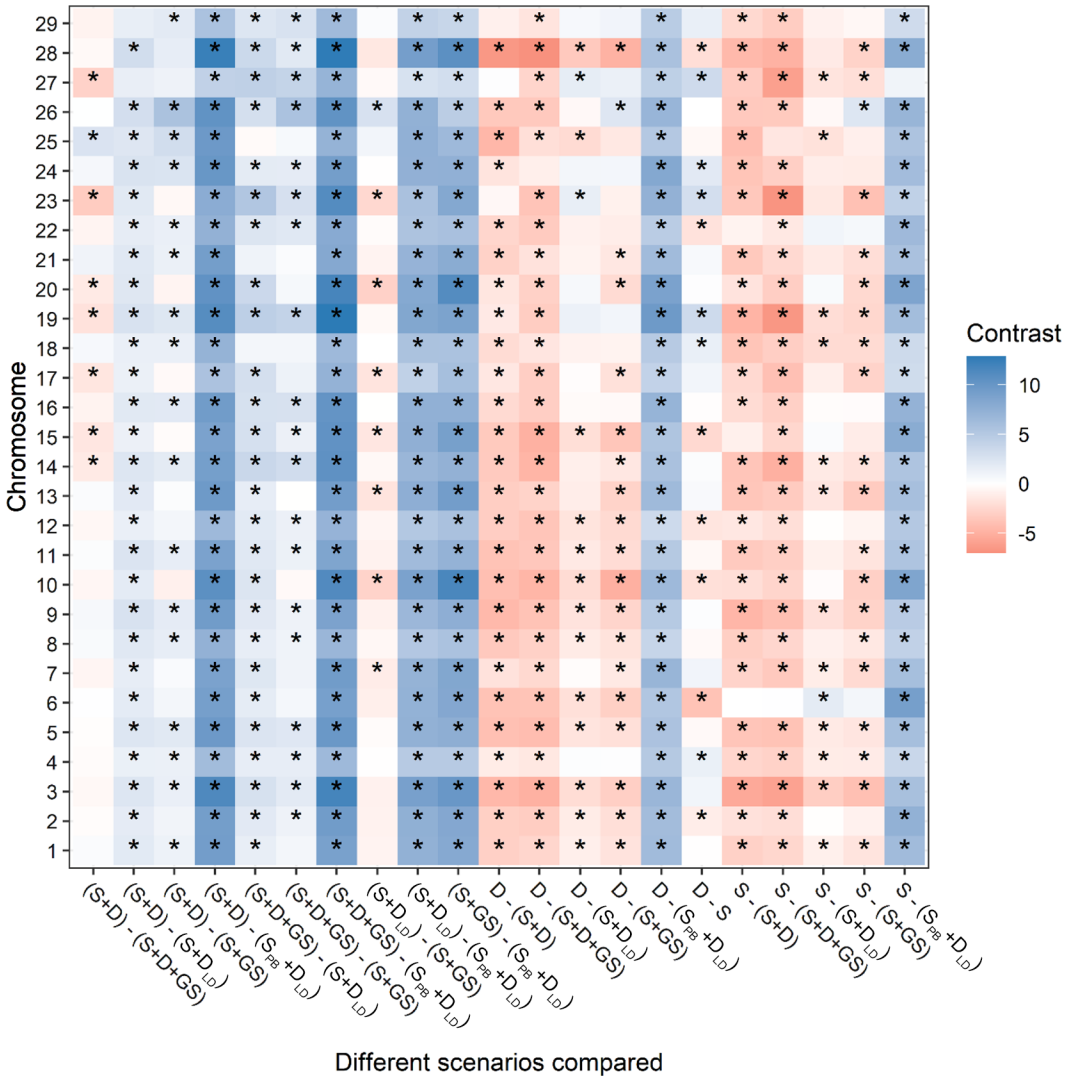

Figure 3. Comparison between LSM of imputation accuracy by chromosome across pairwise scenarios, with their associated P-values (*$P < 0.05$). S+D = sires and dams genotyped with a 50K chip in reference population; $S+D_{LD}$ = sires genotyped with a 50K chip and dams genotyped with a 7K chip in reference population; D = dams genotyped with a 50K chip in reference population; S+GS = sires and maternal grandsires genotyped with a 50K chip in reference population; S+D+GS = sires, dams, and maternal grandsires genotyped with a 50K chip in reference population; $S_{PB}+D_{LD}$ = purebred bulls from 5 different bulls genotyped with a 50K chip and dams genotyped with a 7K chip in the reference population.

of imputation scenario S and S+D, the IA of 5 common chromosomes of 14, 17, 19, 23, and 27 was significantly increased. Thus, the results of estimated IA by chromosomes showed that incorporating first- or second-order relationships of individuals in the validation population into the imputation scenarios did not result in the same set of chromosomes increasing the IA.

DISCUSSION

The present study aimed to evaluate the quality of IA_m and IA_i in US crossbred dairy cattle. It was carried out by exploring the IA_m and IA_i, as well as their distributions, from imputing 7K to 50K SNP chips, the most common chips used by dairy breeders. The effects of the size and relationships between the reference and validation population were also evaluated.

Imputation Accuracy for Crossbred Versus Purebred Dairy Population

When related individuals were included in the reference population scenarios, the estimated IA_m and IA_i for crossbreds ranged from 86.70 to 90.09%, whereas purebreds exceeded 90.75%. Our IA results for crossbreds were consistent with those imputed from 20K

to 50K in a tropical crossbred dairy cattle population, ranging from 50 to 94% (Oliveira Júnior et al., 2017). However, estimated IA in our study were greater than those found in crossbred Canadian beef reported by Ventura et al. (2014) imputing 6K to 50K SNP chip, and they investigated the IA for crossbreds using a reference population of older purebreds, showing that it ranged from 54 to 74% accuracy. This difference in the magnitude of IA could be explained in part by the fact that their population was admixed, whereas the imputed crossbreds in the validation population in our study were 50% Jersey and 50% Holstein. The reference population in our study had a strong relationship (sires, dams, maternal grandsire, or combination of those) with validation population for crossbred, which was the explanation of greater estimated IA compared with the study of Ventura et al. (2014). When only purebreds were used as a reference population to impute genotyped crossbred in a validation population for African dairy cattle, the IA based on correlation between masked and imputed genotypes was reported to be less than 82% using 3 different imputation software programs (Aliloo et al., 2018). However, when related purebred breeds were added to the reference population of imputation scenarios in this study, the estimated IA were greater than those observed by Aliloo et al. (2018).

Numerous studies showed high IA for using purebred dairy in the reference population to impute the same purebred breed, which was confirmed by our findings for Jersey and Holstein cattle. For instance, the percentage of correctly imputed genotypes for a subset of purebred US Holsteins to impute the same population were 90.5% from 3K to 99.0% from 50K using findhap software, and 91.1% from 3K to 99.3% from 50K using FImpute (VanRaden et al., 2013b). In addition, a study to impute purebred US Jersey cows from 2K or 4K to 50K SNP chips revealed IA in the range of 80 to 95% (Weigel et al., 2010b). In our study, the same imputation scenario was applied to compare IA for crossbred versus purebred Jersey and Holstein. Thus, it is concluded that IA in purebred cows was significantly higher than in crossbred cows because the inclusion of 2 different purebred breeds in reference population increases the number of probable haplotypes, and long-range haplotypes in crossbreds will likely differ from those in a purebred reference population (Aliloo et al., 2018), which may explain the lower estimated IA in crossbreds compared with purebreds. A significant correlation between homozygosity and IA was also revealed and highlighted in our study. The higher IA in purebred dairy populations compared with crossbred dairy populations may also be due to homozygous regions, which were more common in purebred cows than in crossbred cows.

Effects of Reference Population Scenarios on Crossbreds in Validation

In this study, the estimated IA in crossbred and purebred populations significantly increased as the proportion of relationship between individuals in the reference and validation population increased. Aliloo et al. (2018) reported similar findings for African crossbred dairy cows and discovered that IA increased as the genomic relationship between the reference and validation populations increased. The importance of relatedness between reference and validation populations was also confirmed by Oliveira Júnior et al. (2017) for increasing IA for tropical crossbred dairy cattle. In a multi-breed beef cattle population, Ventura et al. (2014) reported higher IA when closely related individuals, along with a representation of the breed composition of the imputed group, were included in the reference population. The recommendation to add closely related animals to impute genotypes of crossbred animals has also been made in other species, such as pigs (Xiang et al., 2015) and sheep (Ventura et al., 2016). Indeed, closely related individuals shared longer haplotype segments, which were used to infer missing markers. The imputation methodology used in our study by findhap (version 4) software (VanRaden, 2015) constructed haplotype segments that were as large as possible and iteratively moved to smaller ones if no consistent haplotypes were found in the reference population. Thus, when reference and validation populations were unrelated, they shared very short haplotypes, explaining the lower IA in these scenarios. In addition, this software considers the pedigree information to phase the haplotypes, allowing higher IA when related individuals are present in the reference population.

In this study, it is important to note that all of the data were phased ahead of time using the 78K SNP chip genotypes and Beagle 5.3 software (Browning et al., 2021). However, it is worth mentioning that real-world data may be phased using medium- or low-density chips, which could result in slightly lower imputation accuracies than those observed in this study.

Based on our results, adding 1,000 unrelated purebred animals in the reference population when related individuals are already present in the reference population did not increase IA and can even reduce these values and reduce the percentage of well-imputed markers and individuals. One reason could be that the reference population was less related to the population of crossbred animals in the validation population than in scenarios where only closely related animals are added, based on the GRC (Supplemental Table S4). According to published research, Ventura et al. (2014) showed that adding purebred animals to the reference population of

another purebred population did not improve the IA from a 6K to a 50K SNP chip. In addition, inclusion of purebred Holstein and Gyr in the reference population had limited or no gain in IA for imputing crossbred Girolando (Gyr × Holstein) dairy cattle (Oliveira Júnior et al., 2017). In the literature, the addition of crossbred animals to the reference population improved the IA in Girolando (Oliveira Júnior et al., 2017), tropical crossbreds (Aliloo et al., 2018), and US crossbreds (VanRaden et al., 2020). Based on our analysis, we expected an increase in IA for crossbreds as they shared more common haplotypes with the imputed crossbred population compared with purebreds. However, adding 1,000 crossbred animals to the reference population did not result in improved IA for HO×JE or JE×HO crossbreds in the validation population, as per our study.

The fact that IA have not improved in our study may be due to using genotypes of crossbred animals from different breeds in the reference population, with few genotype data available for HO×JE and JE×HO to impute our Holstein and Jersey crossbreds in the validation population. Additionally, the GRC between the validation and reference populations was lower in the scenario with the 1,000 crossbred animals randomly added to the reference population than it was in the scenarios without these animals (Supplemental Table S4). The crossbred animals in the reference population likely had few haplotypes in common with the crossbred animals in the validation population, and therefore no increase in IA was observed. Another hypothesis suggests that IA has already reached a maximum value for scenarios that included parents (i.e., the closest individuals). In fact, we also noticed results of the same nature in purebred Jersey and Holstein, and the IA was not improved by the inclusion of 1,000 random individuals to the reference population.

Thus, our results showed that it is possible to achieve greater than 85% IA in US crossbred dairy cattle when related individuals of those crossbreds were added to the reference population. The addition of unrelated crossbred genotypes to the reference population had no effect on IA improvement. However, the addition of genotype information from relative crossbred animals to the reference population could not be tested in our study but could increase and improve IA, as already demonstrated by Aliloo et al. (2018).

Effects of Chromosomes on Imputation Accuracy

The estimated IA value for each individual chromosome in our study was similar to the estimated IA for the whole genome. In 3 out of the 7 imputation scenarios, IA value increased as the chromosome length increased, but even for large chromosomes, this difference was minimal. Evidence suggests association between IA and chromosome length in sheep and beef cattle (Sun et al., 2012; Piccoli et al., 2014; Ventura et al., 2016). Chromosomes 1 to 29 were arranged in decreasing order of length, and it was discovered that the last chromosomes frequently had the lowest IA value in different reference population scenarios. In addition, studies on Angus, Braford, and Hereford beef cattle, as well as sheep, had previously supported the finding that the last chromosomes typically had lower IA than the first chromosomes (Sun et al., 2012; Piccoli et al., 2014; Ventura et al., 2016).

In this study, the relatedness between the reference and validation populations did not have the same effect on chromosomal IA, and different chromosomes showed an increase in IA values depending on whether the genomic data from the sire, dams, or maternal grandsire was added to the reference population. However, whenever related animals of validation population included to the reference population of imputation scenarios, the IA values of the first 5 long chromosomes increased noticeably, and the IA values of the last shorter chromosomes decreased. This may be because the shorter the chromosomes, the lower the overall chromosome accuracy, as misattributed distal regions comprise a larger proportion of the overall chromosome (Sun et al., 2012).

To conclude, the IA on each chromosome differed according to the degree of relatedness between reference and validation population for crossbreds. Including more related individuals into reference population seems promising to increase IA, particularly for the first chromosomes but not for the last.

Imputation for US Crossbred Dairy Population

The findings of this study confirmed that IA was in the range of 85 to 90% for HO×JE and JE×HO US crossbred dairy population when imputing from 7K to 50K SNP chip, still lower than in purebred animals. The scenario of S+D+GS achieved the highest imputation accuracy (~90%) for imputing commercial 50K SNP panels. This result is promising because according to the CDCB database, 53% of the HO×JE and JE×HO crossbred animals had available genomic information of the sire, dam, and maternal grandsire (Supplemental Table S1). The second most common scenario in the CDCB database was the S+D_{LD} scenario with IA of around 87%, and the potential addition of related crossbred individuals of HO×JE or JE×HO to the reference population could be beneficial to improve IA for this scenario and should be tested. The third most common scenario in the CDCB database is the scenario S+D (8% JE×HO and HO×JE, Supplemental Table

S1). This case is promising because the IA was only one point lower than in the scenario S+D+GS. However, the scenario S_{PB}+D_{LD}, which is the most frequently used approach by CDCB for genomic evaluation of crossbreds, did not gave the most promising IA results (~80%), so the strategy could eventually be rethought. Thus, adding genomic information from bulls of other breeds than Holstein or Jersey in the reference population would not be a good strategy to impute crossbred animals. In addition, in cases where no genotyped related individuals were available, the best scenario was to put crossbred individuals in the reference population rather than purebred animals.

The breeding scheme for dairy cattle is already an advantage for the imputation of crossbred, as dairy cattle are characterized by a small number of sires with large family sizes and complete pedigree information in many generations. This structure is a considerable advantage for the imputation compared with other species such as beef cattle (Ventura et al., 2014) or sheep (Hayes et al., 2012), where alternative solutions have to be found to face the problem of pedigree structure.

This study determined IA values only for F_1 HO×JE and JE×HO crossbreds. The same type of analysis can also be verified later in other types of crosses commonly found in the United States, such as Ayrshire × Holstein and Brown Swiss × Holstein. The results can also be investigated for crossbreds rather than F_1 crossbreds. Indeed, 19% of crossbreds with genomic data in the CDCB database result from of a cross between a purebred animal and an already crossed individual (Supplemental Table S1), such as backcross, second backcross, or 3-breed backcross, for example. Different approaches will likely need to be considered in this type of situation, and the makeup of the reference panel will need to be examined.

The accuracy of the imputation is crucial for the estimation of correct genetic values. Numerous studies on purebred Jersey and Holstein cows revealed that the accuracy of genomic selection did not appear to decrease with imputation (Weigel et al., 2010a; VanRaden, et al., 2011; Mulder et al., 2012). However, for crossbred dairy populations, no study confirms that the accuracy of genomic selection does not decline with imputation. The study by VanRaden et al. (2020) did not observe the quality of imputation, but highlighted that genomic predictions weighted by BBR were slightly more accurate than predictions using only the predominant breed. Thus, one of the solutions for estimating genetic values of crossbreds could be to consider the BBR to estimate the genetic values, as suggested by VanRaden et al. (2020). Further studies will need to ensure that the accuracy of the genetic values is not affected by the imputation of the data, even if some studies explained that imputations with high error rates and bias from wrongly inferred genotypes will not propagate in accuracy of subsequent genomic predictions (Wu et al., 2016).

CONCLUSIONS

This study aimed to evaluate the IA for the US crossbred dairy cattle population (HO×JE and JE×HO). The results may provide information to assist future studies involving genomic data in crossbred US dairy cattle. The highest IA for crossbreds were shown for the S+D+GS scenario and confirmed the importance of adding relative individuals of crossbred animals in the reference population for imputation strategy. Additional research should be performed on other types of crosses that are also found in the United States, with a focus on ensuring that the influence of imputation on the genetic evaluation of these crossbreds is kept to a minimum.

ACKNOWLEDGMENTS

The authors thank the Council on Dairy Cattle Breeding (CDCB, Bowie, MD) for providing the data in this study. The authors have not stated any conflicts of interest.

REFERENCES

Aliloo, H., R. Mrode, A. M. Okeyo, G. Ni, M. E. Goddard, and J. P. Gibson. 2018. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. J. Dairy Sci. 101:9108–9127. https://doi.org/10.3168/jds.2018-14621.

Browning, B. L., X. Tian, Y. Zhou, and S. R. Browning. 2021. Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. 108:1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005.

Calus, M. P. L., A. C. Bouwman, J. M. Hickey, R. F. Veerkamp, and H. A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: A review of livestock application. Animal 8:1743. https://doi.org/10.1017/S1751731114001803.

Chud, T. C. S., R. V. Ventura, F. S. Schenkel, R. Carvalheiro, M. E. Buzanskas, J. O. Rosa, M. de Alvarenga 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.

Déru, V. 2023. Supplemental Table S1. Figshare. Figure. https://doi.org/10.6084/m9.figshare.21858864.

Déru, V., F. Tiezzi, P. M. VanRaden, E. Lozada-Soto, S. Toghiani, and C. Maltecca. 2023a. Supplemental Table S2 for Imputation accuracy from low- to medium-density SNP chips for US crossbred dairy cattle. Figshare. https://doi.org/10.6084/m9.figshare.21858906.v1.

Déru, V., F. Tiezzi, P. M. VanRaden, E. Lozada-Soto, S. Toghiani, and C. Maltecca. 2023b. Supplemental Table S3 for Imputation accuracy from low- to medium-density SNP chips for US cross-

bred dairy cattle. Figshare. https://doi.org/10.6084/m9.figshare.21858924.v1.
Déru, V., F. Tiezzi, P. M. VanRaden, E. Lozada-Soto, S. Toghiani, and C. Maltecca. 2023c. Supplemental Table S4 for Imputation accuracy from low- to medium-density SNP chips for US crossbred dairy cattle. Figshare. https://doi.org/10.6084/m9.figshare.21858930.
Déru, V., F. Tiezzi, P. M. VanRaden, E. Lozada-Soto, S. Toghiani, and C. Maltecca. 2023d. Supplemental Table S5 for Imputation accuracy from low- to medium-density SNP chips for US crossbred dairy cattle. Figshare. https://doi.org/10.6084/m9.figshare.21858948.
Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. J. Van Der Werf. 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43:72–80. https://doi.org/10.1111/j.1365-2052.2011.02208.x.
Heins, B. J., L. B. Hansen, A. J. Seykora, A. R. Hazel, D. G. Johnson, and J. G. Linn. 2008. Crossbreds of Jersey × Holstein compared with pure Holsteins for body weight, body condition score, dry matter intake, and feed efficiency during the first one hundred fifty days of first lactation. J. Dairy Sci. 91:3716–3722. https://doi.org/10.3168/jds.2008-1094.
Lenth, R. V. 2016. Least-squares means: The R package lsmeans. J. Stat. Softw. 69. https://doi.org/10.18637/jss.v069.i01.
Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876–889. https://doi.org/10.3168/jds.2011-4490.
Norman, H. D., F. L. Guinan, J. H. Megonigal, and J. W. Dürr. 2021. Reasons that cows in dairy herd improvement programs exited the milking herd in 2021. Accessed Nov. 10, 2022. https://queries.uscdcb.com/publish/dhi/current/cullall.html.
Oliveira Júnior, G. A., T. C. S. Chud, R. V. Ventura, D. J. Garrick, J. B. Cole, D. P. Munari, J. B. S. Ferraz, E. Mullart, S. DeNise, S. Smith, and M. V. G. B. da Silva. 2017. Genotype imputation in a tropical crossbred dairy cattle population. J. Dairy Sci. 100:9623–9634. https://doi.org/10.3168/jds.2017-12732.
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.
R Core Team. 2016. R: A language and environment for statistical computing. R. Foundation for Statistical Computing.
Schefers, J. M., and K. A. Weigel. 2012. Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. Anim. Front. 2:4–9. https://doi.org/10.2527/af.2011-0032.
Sørensen, M. K., E. Norberg, J. Pedersen, and L. G. Christensen. 2008. Invited Review: Crossbreeding in dairy cattle: A Danish perspective. J. Dairy Sci. 91:4116–4128. https://doi.org/10.3168/jds.2008-1273.
Sun, C., X.-L. Wu, K. A. Weigel, G. J. M. Rosa, S. Bauck, B. W. Woodward, R. D. Schnabel, J. F. Taylor, and D. Gianola. 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet. Res. (Camb.) 94:133–150. https://doi.org/10.1017/S001667231200033X.
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. 2015. findhap.f90: Find haplotypes and impute genotypes using multiple chip sets and sequence data. Accessed Oct. 29, 2022. https://www.ars.usda.gov/northeast-area/beltsville-md-barc/beltsville-agricultural-research-center/agil/aip/software/findhap/.
VanRaden, P. M., T. A. Cooper, G. R. Wiggans, J. R. O'Connell, and L. R. Bacheller. 2013a. Confirmation and discovery of maternal grandsires and great-grandsires in dairy cattle. J. Dairy Sci. 96:1874–1879. https://doi.org/10.3168/jds.2012-6176.
VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. C. H. M. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, and G. A. Doak. 2013b. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 96:668–678. https://doi.org/10.3168/jds.2012-5702.
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.
VanRaden, P. M., M. E. Tooker, T. C. S. Chud, H. D. Norman, J. H. Megonigal Jr., I. W. Haagen, and G. R. Wiggans. 2020. Genomic predictions for crossbred dairy cattle. J. Dairy Sci. 103:1620–1631. https://doi.org/10.3168/jds.2019-16634.
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.
Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010a. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93:5423–5435. https://doi.org/10.3168/jds.2010-3149.
Weigel, K. A., C. P. Van Tassell, J. R. O'Connell, P. M. VanRaden, and G. R. Wiggans. 2010b. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93:2229–2238. https://doi.org/10.3168/jds.2009-2849.
Wiggans, G. R., T. A. Cooper, P. M. VanRaden, K. M. Olson, and M. E. Tooker. 2012. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. J. Dairy Sci. 95:1552–1558. https://doi.org/10.3168/jds.2011-4985.
Wiggans, G. R., P. M. VanRaden, D. J. Null, E. L. Nicolazzi, G. B. Jansen, and J. H. Megonigal. 2021. Genomic evaluation of crossbred dairy cattle in the United States—An update. Interbull Bull. 56. Accessed Oct. 29, 2022. https://www.ars.usda.gov/ARSUserFiles/80420530/Publications/Scientific/Conferences/2021/IB_56_17-21_WiggansEtAl.pdf.
Wu, X.-L., J. Xu, G. Feng, G. R. Wiggans, J. F. Taylor, J. He, C. Qian, J. Qiu, B. Simpson, J. Walker, and S. Bauck. 2016. Optimal design of low-density SNP arrays for genomic prediction: Algorithm and applications. PLoS One 11:e0161719. https://doi.org/10.1371/journal.pone.0161719.
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.

ORCIDS

Vanille Déru https://orcid.org/0000-0003-1722-5698
Francesco Tiezzi https://orcid.org/0000-0002-4358-9236
Paul M. VanRaden https://orcid.org/0000-0002-9123-7278
Emmanuel A. Lozada-Soto https://orcid.org/0000-0001-8381-7988
Sajjad Toghiani https://orcid.org/0000-0002-5090-728X
Christian Maltecca https://orcid.org/0000-0002-9996-4680