The impact of reducing the frequency of animals genotyped at higher density on imputation and prediction accuracies using ssGBLUP¹

Bruna P. Sollero,^{*,2} Jeremy T. Howard,[†] and Matthew L. Spangler[‡]

*Embrapa Pecuária Sul, Bagé, RS 96401-970, Brazil; [†]Smithfield Premium Genetics, Rose Hill, NC 28458; and [‡]Animal Science Department, University of Nebraska-Lincoln, Lincoln, NE 68527

ABSTRACT: The largest gains in accuracy in a genomic selection program come from genotyping young selection candidates who have not yet produced progeny and who might, or might not, have a phenotypic record recorded. To reduce genotyping costs and to allow for an increased amount of genomic data to be available in a population, young selection candidates may be genotyped with low-density (LD) panels and imputed to a higher density. However, to ensure that a reasonable imputation accuracy persists overtime, some parent animals originally genotyped at LD must be re-genotyped at a higher density. This study investigated the long-term impact of selectively re-genotyping parents with a medium-density (MD) SNP panel on the accuracy of imputation and on the genetic predictions using ssGBLUP in a simulated beef cattle population. Assuming a moderately heritable trait (0.25) and a population undergoing selection, the simulation generated sequence data for a founder population (100 male and 500 female individuals) and 9,000 neutral markers, considered as the MD panel. All selection candidates from generation 8 to 15 were genotyped with LD panels corresponding to a density of 0.5% (LD_0.5), 2% (LD_2), and 5% (LD_5) of

the MD. Re-genotyping scenarios chose parents at random or based on EBV and ranged from 10% of male parents to re-genotyping all male and female parents with MD. Ranges in average imputation accuracy at generation 15 were 0.567 to 0.936, 0.795 to 0.985, and 0.931 to 0.995 for the LD_0.5, LD_2, and LD_5, respectively, and the average EBV accuracies ranged from 0.453 to 0.735, 0.631 to 0.784, and 0.748 to 0.807 for LD_0.5, LD_2, and LD_5, respectively. Re-genotyping parents based on their EBV resulted in higher imputation and EBV accuracies compared to selecting parents at random and these values increased with the size of LD panels. Differences between re-genotyping scenarios decreased when the density of the LD panel increased, suggesting fewer animals needed to be re-genotyped to achieve higher accuracies. In general, imputation and EBV accuracies were greater when more parents were re-genotyped, independent of the proportion of males and females. In practice, the relationship between the density of the LD panel used and the target panel must be considered to determine the number (proportion) of animals that would need to be re-genotyped to enable sufficient imputation accuracy.

Key words: beef cattle, genomic selection, genotyping strategy, imputation, low-density panels

© The Author(s) 2019. Published by Oxford University Press on behalf of the American Society of Animal Science. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

J. Anim. Sci. 2019.97:2780–2792

doi: 10.1093/jas/skz147

¹The first author wishes to thank Embrapa—Empresa Brasileira de Pesquisa Agropecuária (BRAZIL) for financial support. The authors wish to thank the Holland Computing Center-UNL staff for technical assistance.

²Corresponding author: bruna.sollero@embrapa.br

Received February 20, 2019.

Accepted April 25, 2019.

INTRODUCTION

Genomic selection has become pervasive across the majority of livestock species. The ability to increase the accuracy of EBV for selection candidates enables more accurate selection decisions,

particularly in the case whereby traits might not be recorded until later in life or are sex-limited. In the beef and dairy industries, where generation intervals have historically been long and thus the crux of making faster genetic gain, genomic selection enables the use of younger parents without large sacrifices in prediction accuracy. Unfortunately, genotyping represents an increased cost that is additive to already expensive genetic evaluation infrastructures. Although somewhat speculative, it could be reasoned that extensively raised species such as sheep and beef cattle, where there exist numerous owners and decision-makers, are most sensitive to this increased cost.

Profit margins from beef cattle operations may not support the use of a medium- or high-density SNP panel (e.g., Illumina BovineSNP50 BeadChip—50K or 777K). With the advent of low-cost, low-density (LD) genotyping, it may be advantageous to use LD panels as a preliminary screen (Pryce and Daetwyler, 2012) to allow an increased amount of genomic data available or the extension of genomic selection to a larger portion of the population in a cost-efficient way. The introduction of LD SNP panels has made it more attractive to genotype more females in dairy cattle, for example, and may allow opportunities to apply different breeding schemes using genomic technology in beef cattle. Different genomic breeding schemes have been tested, especially in dairy cattle (Moser et al., 2010; Thomasen et al., 2014), in which bulls would be genotyped once or as reported by Winkelman and Spelman (2010), including a first genomic screen to identify an empirical number of candidates (with higher EBV) for the second or full screen. Therefore, selection candidates can be genotyped with lower cost SNP panels and imputed to higher densities (Carvalho et al., 2014; Raoul et al., 2017; Tsai et al., 2017; Lopes et al., 2018).

However, after genotyping selection candidates with low density panels, re-genotyping target animals at higher density over-time is needed to ensure reasonable imputation accuracy persists. Missing genotypes of descendants can be imputed accurately using LD marker sets if ancestor haplotypes are available. Therefore, there is still an opportunity to investigate a scheme or strategy to determine which animals to re-genotype in a beef cattle population based on optimizing cost and impact relative to imputation accuracy. Ideally, there should be enough markers on a LD panel to enable imputation, and then the population could be imputed to a higher density for the purposes of routine genetic evaluation.

In the case of genomic selection, imputation strategies may overcome the problem of relying on

a small reference population genotyped at higher density, which is especially relevant for programs beginning genomic selection or ones with a small number of active animals contained within the breed. The influence of different groups of reference animals on genomic predictions to guide genotyping strategies has been extensively reported (Boligon et al., 2012; Pszczola and Calus, 2016; Eynard et al., 2017; Lopes et al., 2018), but selecting target animals to re-genotype is largely unexplored and is currently performed ad hoc. Due to the increased availability of individual genotypic and phenotypic information, there are more possibilities of designing reference populations with mixed marker panel densities by targeting only relevant individuals to invest in genotyping with higher densities (Eynard et al., 2017).

The application of single-step genomic BLUP (ssGBLUP) has been the method of choice for many organizations across multiple species (Cardoso et al., 2015; Lourenco et al., 2015b; Campos et al., 2018) due to the increase in accuracies of genomic breeding values compared with multiple-step approaches. Simulation studies have compared different methods and addressed genotyping strategies for genomic selection with ssGBLUP (Lourenco et al., 2013; Howard et al., 2018; Piccoli et al., 2018) and imputation strategies using different densities of SNP panels in sheep (Raoul et al., 2017) and dairy cattle (Daetwyler et al., 2011; VanRaden et al., 2011). However, strategies including re-genotyping are largely unexplored. Consequently, the aim of this study was to investigate the long-term impact of selectively re-genotyping parents with a medium-density (MD) SNP panel on the accuracy of imputation and on the genetic predictions using ssGBLUP in a simulated beef cattle population.

MATERIALS AND METHODS

All data were simulated and thus animal care and use approval was not required.

Simulated Data

In order to investigate the long-term impact of selectively re-genotyping parents with a MD SNP panel on the accuracy of imputation and genetic predictions using ssGBLUP for a moderately heritable trait (0.25), 3 autosomal chromosomes, each with a length of 87 Mb representing a partial cattle genome, were simulated using the Geno-Diver software (Howard et al., 2017). To generate linkage disequilibrium levels across the genome that are similar

to a cattle population the “Ne100_Scen1” option within Geno-Diver was utilized when generating sequence data in the founder population. The pattern of linkage disequilibrium decay shown by Howard et al. (2018) is representative of the pattern from the current simulation. After generating sequence data for a founder population containing 100 male and 500 female individuals, 450 quantitative trait loci (QTL) and 9,000 neutral markers, spread equally across the 3 chromosomes, were generated to construct a MD marker panel. The number of markers in the MD panel was chosen to mimic a MD marker panel of Illumina BovineSNP50 BeadChip or GeneSeek Genomic Profiler HD (e.g., 80K). The additive effects were sampled from a Gamma distribution. The additive effects were scaled to generate a trait with a heritability of 0.25. The phenotypic variance was set at 1.0; therefore, the residual variance was 0.75. The forward-in-time simulation approach (Howard et al., 2017) was utilized for 15 generations where the population size was constant. At each generation, the replacement rate of males and females were 0.5 and 0.2, respectively.

Initially, EBV were estimated using PBLUB and after generation 8, were estimated using ssGBLUP (Aguilar et al., 2010; Christensen and Lund, 2010; Aguilar et al., 2011).

The initial genomic relationship matrix (G_{raw}) was constructed as

$$G_{\text{raw}} = \frac{MM'}{2 \sum p_j (1 - p_j)}$$

where M is a genotype incidence matrix that has been centered based on allele frequencies (VanRaden, 2008) and p is the allele frequency of the second allele at the j th SNP. The allele frequencies were estimated from all genotyped animals that were utilized when estimating breeding values. The A_{22} matrix, the portion of the average

numerator relationship matrix (A) corresponding to genotyped animals, was constructed according to Colleau (2002). A weighted genomic relationship ($G_w; 0.95G_{\text{raw}} + 0.05A_{22}$) was utilized when blending genomic and pedigree information and finally to construct the inverse of the hybrid relationship matrix including both pedigree and genomic information (H^{-1}), the τ and Ω values were set at 1.0 to scale the inverse of G_w and A_{22} .

Animals were selected and culled based on EBV, and all animals were culled at age 5. Animals were mated at random, resulting in 1 offspring each with an equal chance to be male or female. The simulated population resulted in 8,100 animals in the 15th generation. All animals had phenotypic information and 4,600 animals were genotyped including the selection candidates born from generation 8 to 15 plus parents of animals born in these generations who themselves were born in a generation previous to generation 8. The remaining animals (3,500) born from generation 0 (founders) to 7, that were not parents of animals born in or after generation 8 were not genotyped. The simulation was replicated 5 times. Figure 1 illustrates the genotyping strategy.

Low-Density Panels and Imputation Strategy

Genotyping of all selection candidates born within a generation (500 per generation) began at generation 8. Selected animals (parents) from generation 8 until generation 15 were genotyped or re-genotyped with the MD panel (9,000 markers) and the remainder of the parents and all selection candidates were genotyped with LD panels corresponding to either 0.5% (50 markers), 2% (180 markers), or 5% (450 markers) of the MD density. Parents (born between generation 3 and 7) of animals born in generation 8 were genotyped with the MD panel and selectively re-genotyping with the MD began in generation 9 (Fig. 1). The LD panels were generated by

Figure 1. Illustration with the genotyping strategy structure of the simulated population from generation 3 to 15. NG blocks represent the proportion of animals (S—sires, D—dams, and/or O—offspring) not genotyped (only phenotyped); MD blocks represent the proportion of animals, for each scenario (parents; S or S+D), genotyped with the medium-density panel; LD blocks represent the animals genotyped with the low-density panels (LD_0.5, LD_2, or LD_5); and RG blocks represent the proportion of animals, according to each scenario (parents; S or S+D), re-genotyped per generation with the medium-density panel. Imputation started from generation 8.

masking the genotypes (equidistant/evenly spaced across the simulated genome) from the MD panel, except for the markers representing the proportion to be evaluated (0.5%, 2%, and 5%).

The FImpute v.2.2 software (Sargolzaei et al., 2014) was used to impute the SNP contained within the MD panel, but not contained within the LD panel, using information from the reference individuals (genotyped with the MD marker panel) based on an overlapping sliding window approach, which captures haplotype similarity between close relatives. Complete pedigree information, beginning at the founders, was used for the analysis. Imputation occurred at each generation before EBV was estimated.

Genotyping Strategies

Different scenarios relative to the proportion of animals to be re-genotyped and strategies to choose animals to be re-genotyped were investigated.

The scenarios varied in terms of the percentage of only males (10%, 20%, 50%, and 80%) or equal proportions of male and female parents (5%, 10%, 25%, 40%, and 50% of each sex) re-genotyped with the MD panel. Animals were chosen either at random or based on their EBV (Top_EBV). A third strategy of choosing animals to re-genotype was initially tested based on the number of offspring (2 scenarios where only male parents with more than 10 or 20 offspring were re-genotyped and 2 other scenarios considering those criteria for male parents plus female parents with more than 5 offspring). Given results (EBV and imputation accuracies) from this strategy were similar to those from the other 2 strategies, results are shown in [Supplementary Table S1](#) but not discussed extensively herein. The similarity of results from choosing animals based on the number of offspring with the other 2 strategies is likely due to the fact that matings were randomly assigned in the simulation and the variation in progeny number between parents was lower than might be expected in field data. Scenarios in which all males (100%) and all males and females (100% and 100%) were re-genotyped, as well as when all animals (parents and selection candidates) were genotyped only with the LD panel, were also tested as “base scenarios” for comparisons.

All scenarios within strategies ([Table 1](#)) were evaluated for each of the 3 LD panels. In total, 75 scenarios were evaluated.

Imputation Accuracies, EBV Accuracies, and Bias of EBV

Beginning at generation 8, the percentage of correctly imputed genotypes (concordance) was

Table 1. Description of each scenario representing the proportion of animals (parents) re-genotyped with the medium-density panel

Scenarios	Description
10_0	10% of male parents re-genotyped with MD panel
5_5	5% of male and female parents re-genotyped with MD panel
20_0	20% of male parents re-genotyped with MD panel
10_10	10% of male and female parents re-genotyped with MD panel
50_0	50% of male parents re-genotyped with MD panel
80_0	80% of male parents re-genotyped with MD panel
25_25	25% of male and female parents re-genotyped with MD panel
40_40	40% of male and female parents re-genotyped with MD panel
50_50	50% of male and female parents re-genotyped with MD panel
100_0	All male parents re-genotyped with MD panel
100_100	All male and female parents re-genotyped with MD panel
No_Imp	No animals genotyped with MD panel and all animals genotyped with LD panel

computed at each generation for imputed SNP to assess imputation accuracy. The imputed markers were compared with the actual markers present in the original MD panel, and thus the proportion of genotypes that were imputed correctly or erroneously was calculated. The average concordance rate of imputed genotypes among the 5 replicates was calculated within each generation and scenario and results were compared among the selection candidates genotyped with the LD panel in the last generation (generation 15).

The prediction accuracies were calculated as the Pearson correlation between true breeding value (TBV) and EBV. The mean prediction accuracies of selection candidates from the last generation were computed over 5 replicates. The degree to which EBV were biased across different scenarios for the selection candidates at generation 15 was quantified by the coefficient of regression of TBV on EBV.

Within each replicate at generation 15, the mean TBV were also calculated (genetic gain).

RESULTS AND DISCUSSION

Imputation Accuracy

Strategies. Mean imputation accuracies are reported in [Table 2](#). At generation 15, these values ranged from 0.567 to 0.936, 0.795 to 0.985, and 0.931 to 0.995 for the LD_0.5, LD_2, and LD_5, respectively.

[Figures 2](#) and [3](#) illustrate the imputation accuracies of 5 scenarios (10_0, 10_10, 50_0, 25_25, 50_50, and 100_100) selected to represent

Table 2. Number of animals genotyped/re-genotyped with MD panel (*N*) until generation 15, mean EBV and imputation accuracies (EBV_Acc and Imp_Acc) and respective standard deviation (SD) across scenarios, strategies (random and Top_EBV), and low-density panels (LDP: LD_0.5, LD_2 and LD_5) tested for selection candidates born in the last generation

LDP	Scenarios	<i>N</i>	Random			Top_EBV			Scenarios			Base ¹		
			EBV_Acc	SD	Imp_Acc	EBV_Acc	SD	Imp_Acc	EBV_Acc	SD	Imp_Acc	EBV_Acc	SD	Imp_Acc
			SD	Imp_Acc	SD	SD	Imp_Acc	SD	SD	Imp_Acc	SD	SD	Imp_Acc	SD
LD_0.5 (50 mks)	10_0	45	0.502	0.051	0.567	0.006	0.477	0.041	0.600	0.015	100_0	0.617	0.031	0.835
	5_5	86	0.453	0.073	0.572	0.008	0.452	0.057	0.613	0.015	100_100	0.735	0.020	0.936
	20_0	90	0.473	0.025	0.601	0.013	0.459	0.035	0.633	0.012	No_Imp	0.434	0.081	
	10_10	165	0.461	0.052	0.593	0.014	0.478	0.078	0.648	0.013				
	50_0	225	0.511	0.043	0.692	0.012	0.525	0.029	0.739	0.010				
	80_0	360	0.579	0.047	0.792	0.005	0.578	0.030	0.803	0.004				
	25_25	416	0.521	0.049	0.706	0.013	0.575	0.048	0.782	0.008				
	40_40	660	0.614	0.050	0.814	0.009	0.618	0.037	0.858	0.008				
	50_50	825	0.666	0.016	0.856	0.007	0.660	0.043	0.889	0.007				
	10_0	45	0.631	0.022	0.795	0.010	0.655	0.009	0.807	0.007	100_0	0.755	0.030	0.953
LD_2 (180 mks)	5_5	86	0.676	0.041	0.811	0.004	0.675	0.034	0.834	0.009	100_100	0.784	0.021	0.985
	20_0	90	0.682	0.036	0.832	0.006	0.654	0.022	0.850	0.007	No_Imp	0.632	0.014	
	10_10	165	0.697	0.024	0.844	0.007	0.705	0.035	0.867	0.008				
	50_0	225	0.712	0.031	0.897	0.003	0.713	0.012	0.914	0.002				
	80_0	360	0.720	0.025	0.938	0.002	0.750	0.042	0.943	0.003				
	25_25	416	0.726	0.029	0.902	0.003	0.755	0.026	0.932	0.002				
	40_40	660	0.756	0.012	0.946	0.006	0.761	0.027	0.959	0.004				
	50_50	825	0.758	0.034	0.958	0.003	0.749	0.027	0.969	0.001				
	10_0	45	0.777	0.026	0.931	0.006	0.773	0.028	0.933	0.004	100_0	0.807	0.016	0.984
	5_5	86	0.748	0.024	0.944	0.004	0.775	0.026	0.946	0.003	100_100	0.806	0.013	0.995
LD_5 (450 mks)	20_0	90	0.776	0.037	0.949	0.003	0.795	0.037	0.952	0.002	No_Imp	0.784	0.027	
	10_10	165	0.794	0.018	0.954	0.003	0.781	0.027	0.960	0.004				
	50_0	225	0.777	0.026	0.968	0.003	0.794	0.007	0.973	0.001				
	80_0	360	0.798	0.023	0.980	0.001	0.784	0.015	0.982	0.002				
	25_25	416	0.791	0.028	0.968	0.002	0.801	0.025	0.979	0.002				
	40_40	660	0.814	0.016	0.982	0.001	0.800	0.024	0.987	0.001				
	50_50	825	0.800	0.019	0.986	0.001	0.799	0.010	0.990	0.001				

¹Base scenarios (as described in Table 1) represented by 100_0, 100_100, and No_Imp.

Figure 2. Mean and standard deviation (error bars) for 5 replicates of imputation accuracies of selection candidates in the last generation represented between strategies (random and Top_EBV) across 6 scenarios (as described in Table 1) within low-density panels (A) low-density panel representing 0.5% of SNPs from the MD panel, B) low-density panel representing 2% of SNPs from the MD panel, and C) low-density panel representing 5% of SNPs from the MD panel). Bars filled with diagonal stripes represent the base scenario (100_100) with the same value for both strategies.

Figure 3. Mean and standard deviation (error bars) for 5 replicates of imputation accuracies of selection candidates in the last generation represented between low-density panels (LD_0.5, LD_2, and LD_5) across 6 scenarios (as described in Table 1) within strategies (A) random strategy, B) Top_EBV strategy). Bars filled with diagonal stripes represent the base scenario (100_100).

intervals between the range of the percentage of parents re-genotyped (from 10_0 to 100_100). Re-genotyping parents based on their EBV (Top_EBV) resulted in higher imputation accuracies compared to selecting parents at random. Although the matings were randomly allocated, individuals were selected and culled based on EBV and animals with a higher EBV were more likely to be selected as parents and have offspring in subsequent generations. Consequently, re-genotyping animals based on EBV led to re-genotyping animals with a

higher degree of connectivity with the entire population. However, differences between strategies decreased with increasing marker density (Fig. 2). Differences between selecting animals at random versus selecting animals to be re-genotyped based on EBV varied from 1.3% to 10.7% across scenarios using the LD_0.5 panel, 1.1% to 3.3% for the LD_2 panel, and 0.2% to 1.1% for the LD_5 panel.

Low-density panels. In general, imputation accuracy increased when the size of LD panels

increased (Fig. 3) regardless of strategy. With 0.5% of SNPs covering the genome (LD_0.5), the mean imputation accuracy ranged from 0.567 to 0.936 across scenarios considering random and Top_EBV strategies (Table 2). For the LD panel with 2% of markers, average imputation accuracies were higher (0.795 to 0.985), whereas for 5% of markers, the range of the average was 0.931 to 0.995. Similarly, Raoul et al. (2017) reported lower imputation accuracies with lower density SNP panels, ranging from 84.5% for the smaller LD panel (250 markers) to 96.5% with 1,000 markers using a simulated sheep population. A clear interaction between the strategy and panel density was observed whereby a reduced number of animals selected based on EBV could achieve similar imputation accuracies as randomly selecting a large proportion of animals when the panel density increased.

Daetwyler et al. (2011) simulated a dairy cattle population whereby 3, 2, or 1 parental generations were genotyped with a dense panel (15 SNP/cM) and a proportion of animals in the last generation with sparse genotypes (panels of 420, 1,020, and 3,000 SNP for a 30-M genome) to be imputed. The authors concluded that the proportion of correctly imputed loci increased as the density of the sparse genotypes increased (87%, 94%, and 98% of imputation accuracy, respectively, when using the maximum number of ancestral generations). The results reported herein showed imputation accuracy greater than 0.90 for scenarios with more than 225 animals genotyped with the LD_2 (50% of male parents).

Carvalho et al. (2014) reported averages of imputation accuracies that ranged from 0.89 to 0.98 among different validation sets of lower density SNP chips (ranging from 7K to 75K) using a reference set of 793 Nelore animals genotyped with the Illumina BovineHD chip (HD). They also concluded that for this beef breed, the use of LD panels that contain approximately 15K SNP is cost-effective to genotype economically marginal animals and impute the missing genotypes to the HD panel. This LD density is proportional to the simulated LD_2 panel investigated in the current study.

Zhang et al. (2010) tested 5 different sizes of genotyping arrays, from 384 to 6,000 SNP. The overall allelic imputation error rate obtained ranged from 11.7% (384 SNP) to 2.0% (6,000 SNP). The authors concluded that the benefit of adding more markers will be limited at higher marker densities and that the imputation accuracy was greater when more ancestors were genotyped with the high density panel; in agreement with the inference made from the current study.

Scenarios. Differences between re-genotyping scenarios decreased when the density of the LD panel increased, suggesting fewer animals needed to be re-genotyped to achieve higher imputation accuracies. Imputation accuracy was greater when more parents were re-genotyped, independently of the proportion of males and females re-genotyped, but the inclusion of females seemed to provide a marginal gain particularly when the LD panel density was greater. For example, scenarios where only 10% of males (45 animals) were randomly selected to be re-genotyped (LD_2) resulted in an imputation accuracy of 80%, while using 5% of males and 5% of females (86 animals) resulted in a marginal increase (81%). Re-genotyping only 20% of male parents selected at random (90 animals) resulted in a mean imputation accuracy of 83% (Table 2). The same pattern was observed using the LD_5 panel, but with higher overall imputation accuracies. Chud et al. (2015) suggested that the imputation of females genotyped with LD panels could be carried out using only the males in a reference population of composite beef cattle breed, and that it could be an appropriate strategy for large-scale female selection. Zhang and Druet (2010) evaluated marker imputation with LD panels in Dutch Holstein cattle and assumed that the imputation accuracy for paternal haplotypes was higher than for maternal haplotypes because normally sires are most often genotyped. Overall, if more ancestors are re-genotyped, there is a greater chance to estimate more reliable haplotypes and therefore improve imputation accuracy.

EBV Accuracies

Strategies. When parents were chosen to be re-genotyped based on Top_EBV, accuracies of EBV were slightly increased in most of the scenarios for LD_0.5 and LD_2 compared to randomly choosing animals to re-genotype (Table 2 and Fig. 4). The average EBV accuracies in the last generation of selection candidates within the Top_EBV strategy across scenarios and LD panels without considering the base scenarios ranged from 0.452 to 0.660 (LD_0.5), 0.655 to 0.761 (LD_2), and 0.773 to 0.801 (LD_5), while for the random strategy values varied from 0.453 to 0.666 (LD_0.5), 0.631 to 0.758 (LD_2), and from 0.748 to 0.814 (LD_5). The differences between these 2 strategies might have been larger if the population was not randomly mated. Howard et al. (2018) reported that randomly genotyping resulted in a higher accuracy than selecting animals to genotype based on an index value, even though

Figure 4. Mean and standard deviation (error bars) for 5 replicates of EBV accuracies of selection candidates in the last generation represented between strategies (Random and Top_EBV) across 6 scenarios (as described in Table 1) within low-density panels (A) low-density panel representing 0.5% of SNPs from the MD panel, B) low-density panel representing 2% of SNPs from the MD panel, and C) low-density panel representing 5% of SNPs from the MD panel). Bars filled with diagonal stripes represent the base scenario (100_100).

selectively genotyping based on index value led to an increased rate of genetic change (change in index value). The authors used a simulated beef and swine population undergoing selection, and a key difference compared to the structure of the population investigated in the current study was that only a proportion of animals were genotyped.

The regression coefficients of TBV on EBV (Table 3) across different strategies were on average higher and/or closer to 1 for the scenarios within Top_EBV strategy compared to the random strategy, not considering the base (0.838 vs. 0.832, LD_0.5; 0.963 vs. 0.938, LD_2, and 1.002 vs. 1.007, LD_5, respectively).

Low-density panels. Mean EBV accuracies across scenarios and strategies increased as the density of the LD panel increased (Fig. 5). Following the pattern observed for imputation accuracies, the differences between re-genotyping scenarios decreased as the density of the LD panel increased suggesting fewer animals/parents needed to be re-genotyped to increase EBV accuracy when the LD panel was denser.

When the LD panel was sufficiently dense, re-genotyping and imputation strategies might not be necessary under a ssGBLUP framework as evidenced by the EBV accuracies achieved under the base scenario whereby no animals were re-genotyped using the LD_5 panel (Table 2). On the other hand,

the highest EBV accuracies were observed when all males (from 0.617 to 0.807 across LD panels) or all parents (from 0.735 to 0.806) were re-genotyped.

Scenarios. As illustrated in Figs. 4 and 5, in general, the EBV accuracy increased as the number of animals re-genotyped increased.

The improvement in prediction accuracies when including each gender (male and female) in reference populations (genotyped and phenotyped animals) with different strategies applied for genotyping animals has been addressed for dairy cattle and other species (Pryce and Daetwyler, 2012; Lourenco et al., 2015a; Raoul et al., 2017). Results from the application of ssBLUP in broiler chickens suggested that genotyping both sexes may be a suitable option because of the high reproductive impact that females of this species have (Lourenco et al., 2015a). Similar results were reported by Raoul et al. (2017) using sheep. The authors investigated the impact of including dams either genotyped with MD or LD, and then imputed to MD. Given the contribution of dams to the genetic improvement in sheep, this resulted in a substantial increase in EBV accuracy of selection candidates in the last generation.

However, in beef cattle production systems males have a higher reproductive impact and thus the inclusion of genotyped females may have only a marginal impact relative to imputation accuracy and rates of genetic gain; assuming traits being

Table 3. Number of animals genotyped/re-genotyped with MD panel (N) until generation 15, regression coefficient of true breeding value on estimated breeding value ($b_{TBV,EBV}$) and standard deviation (SD) across scenarios, strategies, and low-density panels (LDP) for selection candidates born in the last generation

LDP	Scenarios ¹	N	$b_{TBV,EBV}$	SD	$b_{TBV,EBV}$	SD	Scenarios ¹	N	$b_{TBV,EBV}$	SD
			Random		Top_EBV				Base	
LD_0.5 (50 mks)	10_0	45	0.878	0.152	0.834	0.154	100_0	450	0.877	0.112
	5_5	86	0.767	0.152	0.755	0.150	100_100	1650	0.966	0.089
	20_0	90	0.764	0.143	0.769	0.149	No_Imp		1.152	0.242
	10_10	165	0.746	0.144	0.794	0.147				
	50_0	225	0.825	0.139	0.803	0.130				
	80_0	360	0.874	0.123	0.858	0.121				
	25_25	416	0.794	0.131	0.899	0.128				
	40_40	660	0.881	0.113	0.904	0.115				
	50_50	825	0.957	0.107	0.927	0.106				
LD_2 (180 mks)	10_0	45	0.881	0.109	0.904	0.105	100_0	450	0.984	0.085
	5_5	86	0.915	0.100	0.938	0.103	100_100	1650	0.995	0.079
	20_0	90	0.918	0.099	0.928	0.107	No_Imp		1.011	0.124
	10_10	165	0.926	0.095	0.989	0.099				
	50_0	225	0.965	0.095	0.957	0.094				
	80_0	360	0.963	0.093	0.993	0.087				
	25_25	416	0.933	0.088	0.982	0.085				
	40_40	660	0.975	0.085	1.006	0.086				
	50_50	825	0.974	0.084	0.982	0.087				
LD_5 (450 mks)	10_0	45	0.996	0.081	0.984	0.081	100_0	450	1.026	0.075
	5_5	86	0.964	0.086	0.990	0.081	100_100	1650	1.037	0.076
	20_0	90	0.992	0.080	0.993	0.076	No_Imp		1.053	0.083
	10_10	165	1.024	0.078	1.001	0.080				
	50_0	225	1.016	0.082	1.023	0.078				
	80_0	360	1.039	0.078	0.988	0.078				
	25_25	416	1.022	0.079	1.018	0.076				
	40_40	660	1.003	0.072	1.009	0.076				
	50_50	825	1.017	0.076	1.013	0.076				

¹Scenarios (as described in Table 1).

Figure 5. Mean and standard deviation (error bars) for 5 replicates of EBV accuracies of selection candidates in the last generation represented between low-density panels (LD_0.5, LD_2, and LD_5) across 6 scenarios (as described in Table 1) within strategies (A) Random strategy, (B) Top_EBV strategy). Bars filled with diagonal stripes represent the base scenario (100_100).

evaluated are not sex-limited. As stated by Pszczola et al. (2012) the highest accuracy was achieved when a large portion of the population was genotyped, and consequently genotyping all animals will always yield the highest population-wide EBV accuracy compared to focusing genotyping efforts only on a subgroup of animals. Thomasen et al.

(2014) simulated genomic data in dairy cattle using ssGBLUP and suggested that compared with a mix of genotyped young bulls and progeny-tested bulls in the reference population, the inclusion of genotyped first-lactation cows resulted in slightly increased annual monetary genetic gain and the reliability of genomic predictions. On the other

hand, comparing gains in predictivity in Angus beef cattle, Lourenco et al. (2015b) found that the addition of top cows to the set with top bulls did not increase the prediction accuracy for any trait using ssGBLUP. In general, results from the current study suggested that mean EBV accuracy was slightly higher when only males were re-genotyped when using either LD_0.5 and LD_5 (Table 2).

The regression coefficients in different scenarios ranged from 0.746 (10_10 with LD_0.5) to 1.152 (No_Imp with LD_0.5) (Table 3). In the current study, bias of the estimated EBV was smaller (closer to 1) when the density of the LD panel increased (LD_2 or LD_5). Edel et al. (2017) stated that selective genotyping and selective imputation in ssGBLUP can contribute to inflated predictions. One possible reason for regression coefficients of EBV being lower than 1 (mostly for scenarios tested with LD_0.5) could be the incomplete LD with causal genes due to the very low coverage and the inability to fully capture additive relationships.

Genetic gain. As expected, genetic merit tended to increase when increasing the proportion of parents re-genotyped with the MD panel (Fig. 6), resulting in greater mean response in TBV of all animals born in the last generation (Table 4). Results for the Top_EBV strategy were mostly higher than

the random strategy specially for LD_0.5. The differences between these values among scenarios decreased as the density of the LD panels increased. This trend suggests that as the density of the LD panel approaches the size of the MD panel, gains from using a mixed density of panels and imputing to a higher density decline. Consequently, once the LD panel becomes sufficiently dense, the LD panel itself is sufficient to conduct genetic evaluation.

General Discussion

The feasibility of using different strategies for selective re-genotyping to enable imputation is critical to develop sound decisions relative to the investment in genomic prediction. From an economic stand point it is important to determine which selection candidates will or will not become an influential parent to then invest in re-genotyping. In the current study, 3 factors affecting imputation accuracy and consequently EBV accuracy were investigated: density of markers on the LD panel, number of animals re-genotyped, and the strategy of selecting parents to be re-genotyped.

The main advantage of using genomic information in animal breeding is to generate more accurate EBV for young animals and those predictions are largely influenced by imputation accuracies and the

Figure 6. Genetic trend represented by the mean of TBV (true breeding value) and standard deviation (error bars) of animals born in the last generation across 6 scenarios (as described in Table 1) and the 2 strategies of re-genotyping (Random and Top_EBV) for the 3 low-density panels tested (A) LD_0.5, B) LD_2, and C) LD_5). Base scenario (100_100) with the same value for both strategies.

Table 4. Mean and standard deviation (SD) of the genetic gain in true breeding value (GG) of selection candidates born at generation 15 across scenarios and strategies (Random and Top_EBV) for the 3 low-density panels (LDP) tested

LDP ¹	Scenarios ²	GG	SD	GG	SD	Scenarios	GG	SD
		Random		Top_EBV			Base ³	
LD_0.5 (50 mks)	10_0	1.74	0.064	1.76	0.093	100_0	1.92	0.044
	5_5	1.72	0.096	1.75	0.069	100_100	2.11	0.081
	20_0	1.77	0.074	1.73	0.095	No_Imp	1.64	0.102
	10_10	1.72	0.079	1.75	0.080			
	50_0	1.80	0.095	1.84	0.038			
	80_0	1.86	0.082	1.83	0.075			
	25_25	1.79	0.102	1.86	0.066			
	40_40	1.93	0.079	1.90	0.056			
	50_50	1.97	0.051	1.98	0.062			
LD_2 (180 mks)	10_0	2.00	0.094	2.01	0.080	100_0	2.09	0.057
	5_5	2.02	0.072	2.01	0.081	100_100	2.20	0.054
	20_0	2.00	0.043	2.01	0.085	No_Imp	1.99	0.057
	10_10	2.07	0.086	2.06	0.092			
	50_0	2.04	0.126	2.08	0.075			
	80_0	2.09	0.087	2.09	0.104			
	25_25	2.11	0.055	2.10	0.090			
	40_40	2.14	0.062	2.09	0.095			
	50_50	2.13	0.097	2.14	0.050			
LD_5 (450 mks)	10_0	2.14	0.096	2.15	0.061	100_0	2.18	0.055
	5_5	2.10	0.055	2.16	0.128	100_100	2.15	0.044
	20_0	2.13	0.056	2.14	0.107	No_Imp	2.11	0.092
	10_10	2.14	0.075	2.19	0.070			
	50_0	2.16	0.079	2.15	0.094			
	80_0	2.16	0.092	2.16	0.051			
	25_25	2.15	0.053	2.15	0.074			
	40_40	2.18	0.074	2.17	0.055			
	50_50	2.18	0.099	2.17	0.069			

¹LDP—low-density panel with 0.5% (LD_0.5 = 50 markers), 2% (LD_2 = 180 markers), and 5% (LD_5 = 450 markers) of SNPs out of the medium-density (MD) panel (9,000 markers).

²Scenarios—as described in Table 1.

³Base scenarios (as described in Table 1) represented by 100_0, 100_100, and No_Imp.

number of genotyped ancestor generations (e.g., Mulder et al., 2012). Genotyped multigeneration pedigrees are suitable to resolving the phase of whole-chromosome haplotypes and therefore result in more accurate genotype imputation and predictions (Tsai et al., 2017). Based on this, the current study considered 8 generations (from 8 to 15) of LD genotyped and re-genotyped parents cumulatively.

As stated by Zhang and Druet (2010), economic and technical considerations must be taken into account to decide whether larger SNP panels should be used. In our case, there was no benefit in applying the LD_5 with the current strategies of re-genotyping. Given ssGBLUP is less sensitive to scenarios where selection candidates are selectively genotyped (Howard et al., 2018), the use of very LD panels and re-genotyping (conditional on economic resources available for genotyping) of male parents

could be advisable to achieve reasonable imputation and prediction accuracies. This may result in decreased cost and sufficient prediction accuracy of selection candidates compared to when applying the same method (ssGBLUP) to all animals genotyped with medium- or high-density panels as tested/reported in other studies (e.g., Silva et al., 2016).

Our results suggested that the best strategy to decide which parents should be re-genotyped should be based on EBV (Top_EBV) and biased toward males. However, schemes of re-genotyping a percentage of male parent animals has to be tested in real data to determine the cost-efficiency. Moreover, the relationship between the LD panel density and the density of the target panel must also be contemplated to ensure sufficient imputation accuracy. Given some breeders may be resistant to re-genotype most animals (VanRaden,

2010), strategically choosing animals to reduce overall cost would be appealing. Admittedly, breed organizations may have to subsidize re-genotyping to benefit the entire population.

CONCLUSIONS

Overall, results showed that greater imputation accuracies were obtained when re-genotyping a larger fraction of parents regardless the size of the LD used. However, given choices must be made under economic constraints, a bias toward re-genotyping males seems logical under normal reproductive scenarios where males contribute more substantially to gene flow of the next generation. However, in the case of embryo transfer, re-genotyping females may become more important. As the panel densities increased, so did imputation accuracies. Although sufficiently dense “LD” panels may be adequate for use in a ssGBLUP framework without the need for re-genotyping, a strategy that optimizes monetary investment and utilizes very LD panels for young animals and identifies only a sample of parents instead of all parents to be re-genotyped might be optimal.

SUPPLEMENTARY DATA

Supplementary data are available at *Journal of Animal Science* online.

LITERATURE CITED

- Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. *J. Dairy Sci.* 93:743–752. doi:10.3168/jds.2009-2730
- Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet.* 128:422–428. doi:10.1111/j.1439-0388.2010.00912.x
- Boligon, A. A., N. Long, L. G. Albuquerque, K. A. Weigel, D. Gianola, and G. J. Rosa. 2012. Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection. *J. Anim. Sci.* 90:4716–4722. doi:10.2527/jas.2012-4857
- Campos, G. S., F. A. Reimann, L. L. Cardoso, C. E. R. Ferreira, V. S. Junqueira, P. I. Schmidt, J. Braccini Neto, M. J. I. Yokoo, B. P. Sollero, A. A. Boligon, et al. 2018. Genomic prediction using different estimation methodology, blending and cross-validation techniques for growth traits and visual scores in Hereford and Braford cattle. *J. Anim. Sci.* 96:2579–2595. doi:10.1093/jas/sky175
- Cardoso, F. F., C. C. Gomes, B. P. Sollero, M. M. Oliveira, V. M. Roso, M. L. Piccoli, R. H. Higa, M. J. Yokoo, A. R. Caetano, and I. Aguilar. 2015. Genomic prediction for tick resistance in Braford and Hereford cattle. *J. Anim. Sci.* 93:2693–2705. doi:10.2527/jas.2014-8832
- Carvalho, R., S. A. Boison, H. H. Neves, M. Sargolzaei, F. S. Schenkel, Y. T. Utsunomiya, A. M. O'Brien, J. Sölkner, J. C. McEwan, C. P. Van Tassell, et al. 2014. Accuracy of genotype imputation in Nelore cattle. *Genet. Sel. Evol.* 46:69. doi:10.1186/s12711-014-0069-1
- Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42:2. doi:10.1186/1297-9686-42-2
- Chud, T. C. S., R. V. Ventura, F. S. Schenkel, R. Carvalho, M. E. Buzanskas, J. O. Rosa, M. de Alvarenga Mudadu, M. V. G. B. da Silva, F. B. Mokry, C. R. Marcondes, et al. 2015. Strategies for genotype imputation in composite beef cattle. *BMC Genet.* 16:1–10. doi:10.1186/s12863-015-0251-7
- Colleau, J. J. 2002. An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol.* 34:409–421. doi:10.1051/gse:2002015
- Daetwyler, H. D., G. R. Wiggans, B. J. Hayes, J. A. Woolliams, and M. E. Goddard. 2011. Imputation of missing genotypes from sparse to high density using long-range phasing. *Genetics* 189:317–327. doi:10.1534/genetics.111.128082
- Edel, C., E. C. G. Pimentel, L. Plieschke, R. Emmerling, and K. Götz. 2017. Effects of selective genotyping and selective imputation in single-step GBLUP. *Interbull.* <https://journal.interbull.org/index.php/ib/article/view/1420> (Accessed 15 April 2017.)
- Eynard, S. E., P. Croiseau, D. Laloë, S. Fritz, M. P. L. Calus, and G. Restoux. 2017. Which individuals to choose to update the reference population? Minimizing the loss of genetic diversity in animal genomic selection programs. *G3: Genes|Genomes|Genetics* 8:g3.1117.2017. doi:10.1534/g3.117.1117
- Howard, J. T., T. A. Rathje, C. E. Bruns, D. F. Wilson-Wells, S. D. Kachman, and M. L. Spangler. 2018. The impact of selective genotyping on the response to selection using single-step genomic best linear unbiased prediction. *J. Anim. Sci.* 96:4532–4542. doi:10.1093/jas/sky330
- Howard, J. T., F. Tiezzi, J. E. Pryce, and C. Maltecca. 2017. Geno-Diver: a combined coalescence and forward-in-time simulator for populations undergoing selection for complex traits. *J. Anim. Breed. Genet.* 134:553–563. doi:10.1111/jbg.12277
- Lopes, F. B., X. L. Wu, H. Li, J. Xu, T. Perkins, J. Genho, R. Ferretti, R. G. Tait, S. Bauck, and G. J. M. Rosa. 2018. Improving accuracy of genomic prediction in Brangus cattle by adding animals with imputed low-density SNP genotypes. *J. Anim. Breed. Genet.* 135:14–27. doi:10.1111/jbg.12312
- Lourenco, D. A., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015a. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genet. Sel. Evol.* 47:56. doi:10.1186/s12711-015-0137-1
- Lourenco, D. A. L., I. Misztal, H. Wang, I. Aguilar, S. Tsuruta, and J. K. Bertrand. 2013. Prediction accuracy for a simulated maternally affected trait of beef cattle using different genomic evaluation models. *J. Anim. Sci.* 91:4090–4098. doi:10.2527/jas.2012-5826

- Lourenco, D. A., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, et al. 2015b. Genetic evaluation using single-step genomic best linear unbiased predictor in American angus. *J. Anim. Sci.* 93:2653–2662. doi:10.2527/jas.2014-8836
- Moser, G., M. S. Khatkar, B. J. Hayes, and H. W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. *Genet. Sel. Evol.* 42:37. doi:10.1186/1297-9686-42-37
- 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. doi:10.3168/jds.2011-4490
- Piccoli, M. L., L. F. Brito, J. Braccini, F. V. Brito, F. F. Cardoso, J. A. Cobuci, M. Sargolzaei, and F. S. Schenkel. 2018. A comprehensive comparison between single- and two-step GBLUP methods in a simulated beef cattle population. *Can. J. Anim. Sci.* 98:565–575. doi:10.1139/cjas-2017-0176
- Pryce, J. E., and H. D. Daetwyler. 2012. Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Anim. Prod. Sci.* 52:107–114. doi:10.1071/AN11098
- Pszczola, M., and M. P. Calus. 2016. Updating the reference population to achieve constant genomic prediction reliability across generations. *Animal* 10:1018–1024. doi:10.1017/S1751731115002785
- Pszczola, M., T. Strabel, J. A. M. van Arendonk, and M. P. L. Calus. 2012. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. *J. Dairy Sci.* 95:5412–5421. doi:10.3168/jds.2012-5550
- Raoul, J., A. A. Swan, and J. M. Elsen. 2017. Using a very low-density SNP panel for genomic selection in a breeding program for sheep. *Genet. Sel. Evol.* 49:76. doi:10.1186/s12711-017-0351-0
- 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. doi:10.1186/1471-2164-15-478
- Silva, R. M., B. O. Fragomeni, D. A. Lourenco, A. F. Magalhães, N. Irano, R. Carvalheiro, R. C. Canesin, M. E. Mercadante, A. A. Boligon, F. S. Baldi, et al. 2016. Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. *J. Anim. Sci.* 94:3613–3623. doi:10.2527/jas.2016-0401
- Thomassen, J. R., A. C. Sørensen, M. S. Lund, and B. Guldbbrandsen. 2014. Adding cows to the reference population makes a small dairy population competitive. *J. Dairy Sci.* 97:5822–5832. doi:10.3168/jds.2014-7906
- Tsai, H.-Y., O. Matika, S. M. Edwards, R. Antolín-Sánchez, A. Hamilton, D. R. Guy, A. E. Tinch, K. Gharbi, M. J. Stear, J. B. Taggart, et al. 2017. Genotype imputation to improve the cost-efficiency of genomic selection in farmed Atlantic salmon. *G3: Genes|Genomes|Genetics* 7:1377–1383. doi:10.1534/g3.117.040717
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. doi:10.3168/jds.2007-0980
- VanRaden, P. M. 2010. Genomic evaluations with many more genotypes and phenotypes. 9th World Congr. Genet. Appl. to Livest. Prod., 1–6 August 2010, Leipzig, Germany. p. 149. doi:10.1046/j.0269-2813.2003.01784.x
- 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. doi:10.1186/1297-9686-43-10
- Winkelman, A. M., and R. J. Spelman. 2010. Response using genome-wide selection in dairy cattle breeding schemes. 9th World Congr. Genet. Appl. to Livest. Prod., 1–6 August 2010, Leipzig, Germany. p. 0290. CD-ROM Communication 714. Available at <http://www.kongressband.de/wcgalp2010/assets/pdf/0290.pdf>.
- Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. *J. Dairy Sci.* 93:5487–5494. doi:10.3168/jds.2010-3501
- Zhang, Z., M. Georges, and T. Druet. 2010. Prediction of missing markers with low density marker panels in dairy cattle. *Proc. WCGALP 2010*, Leipzig, Germany. p. 1–4.