Assessing single-nucleotide polymorphism selection methods for the development of a low-density panel optimized for imputation in South African Drakensberger beef cattle

Simon F Lashmar [1,*], Donagh P Berry [1 2], Rian Pierneef [3], Farai C Muchadeyi [3], Carina Visser [1]

[1] Department of Animal Sciences, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa.

[2] Animal and Grassland Research and Innovation Centre, Teagasc, Moorepark, Fermoy, Co. Cork, Ireland.

[3] Biotechnology Platform, Agricultural Research Council, Private Bag X5, Onderstepoort 0110, South Africa.

* Corresponding author: simon.lashmar@up.ac.za

Abstract

A major obstacle in applying genomic selection (GS) to uniquely adapted local breeds in less-developed countries has been the cost of genotyping at high densities of single-nucleotide polymorphisms (SNP). Cost reduction can be achieved by imputing genotypes from lower to higher densities. Locally adapted breeds tend to be admixed and exhibit a high degree of genomic heterogeneity thus necessitating the optimization of SNP selection for downstream imputation. The aim of this study was to quantify the achievable imputation accuracy for a sample of 1,135 South African (SA) Drakensberger cattle using several custom-derived lower-density panels varying in both SNP density and how the SNP were selected. From a pool of 120,608 genotyped SNP, subsets of SNP were chosen (1) at random, (2) with even genomic dispersion, (3) by maximizing the mean minor allele frequency (MAF), (4) using a combined score of MAF and linkage disequilibrium (LD), (5) using a partitioning-around-medoids (PAM) algorithm, and finally (6) using a hierarchical LD-based clustering algorithm. Imputation accuracy to higher density improved as SNP density increased; animal-wise imputation accuracy defined as the within-animal correlation between the imputed and actual alleles ranged from 0.625 to 0.990 when 2,500 randomly selected SNP were chosen vs. a range of 0.918 to 0.999 when 50,000 randomly selected SNP were used. At a panel density of 10,000 SNP, the mean (standard deviation) animal-wise allele concordance rate was 0.976 (0.018) vs. 0.982 (0.014) when the worst (i.e., random) as opposed to the best (i.e., combination of MAF and LD) SNP selection strategy was employed. A difference of 0.071 units was observed between the mean correlation-based accuracy of imputed SNP categorized as low ($0.01 < MAF \leq 0.1$) vs. high MAF ($0.4 < MAF \leq 0.5$). Greater mean imputation accuracy was achieved for SNP located on autosomal extremes when these regions were populated with more SNP. The presented results suggested that genotype imputation can be a practical cost-saving strategy for indigenous breeds such as the SA Drakensberger. Based on the results, a genotyping panel consisting of ~10,000 SNP selected based on a combination of MAF and LD would suffice in achieving a <3% imputation error rate for a breed characterized by genomic admixture on the condition that these SNP are selected based on breed-specific selection criteria.

Keywords: Sanga cattle; genomics; imputation accuracy; single-nucleotide polymorphisms.

Abbreviations

ACR	allele concordance rate
BTA	*Bos taurus* autosome
DISTMAF	equidistant selection that maximized minor allele frequency
GCR	genotype concordance rate
GS	genomic selection
LD	linkage disequilibrium
LDCLUST	intra-autosomal LD-based clustering algorithm
MAF	minor allele frequency
MAFLD	segment-based selection combining minor allele frequency and linkage disequilibrium
MID	mid-point, equidistant selection
PAM	partitioning-around-medoids
RAN	random selection
SA	South Africa(n)
SNP	single-nucleotide polymorphism

Introduction

Large-scale cattle genotyping is presently undertaken using single-nucleotide polymorphism (SNP)-chips with a vast array of different panels available that vary in both genotype density and breed representation (Nicolazzi et al., 2015). Most of the genotype panels are constructed by selecting SNP that are informative in the most populous breeds of either taurine- (e.g., Bovine SNP50; Matukumalli et al., 2009) or indicine descent (e.g., GeneSeek GGP Indicus; Ferraz et al., 2018). The utility of panels favoring either of these subspecies may therefore not be optimal for less common breeds that often have admixed genomes. This is particularly true as many exploit linkage disequilibrium (LD) between candidate SNP in their generation of lower density panels; LD erodes more rapidly in admixed populations (Shifman, 2003; Toosi et al., 2010) influencing not only the choice of selected SNP but also the ideal SNP density to sufficiently capture the haplotypes segregating in that population.

The South African (SA) Drakensberger is a beef breed with a sleek, black coat that belongs to the Sanga subspecies (*Bos taurus africanus*) of cattle (SA Drakensberger Breeders' Society, 2011) and is indigenous to southern Africa. A study by Makina et al. (2016) that used genome-wide SNP data to investigate patterns of admixture that might assist in elucidating the historic origin of Sanga breeds indicated common ancestry with European and African *Bos taurus* as well as *Bos indicus* breeds. Makina et al. (2016) reported the estimated genomic composition of the SA Drakensberger to be 46% European taurine, 38% African taurine, and 15% indicine; the estimated genomic composition of other Sanga breeds varied slightly with the Afrikaner and Nguni breeds displaying 12 and 15 percentage units greater indicine germplasm, respectively, than the SA Drakensberger. The success of genome-enabled breed improvement strategies, predominantly genomic selection (GS), for breeds such as the SA Drakensberger is therefore contingent on the utility of the SNP included in commercial high-density genotype panels.

The cost of genotyping is still a major barrier to adoption. One potential strategy to reduce the cost of genotyping is to minimize the number of SNP to be genotyped and subsequently impute these SNP to higher density. The SNP selected for a reduced density panel must, however, be (1) abundant enough to facilitate acceptable imputation accuracy to higher density and (2) informative for the breed(s) in which they will be applied. When current genotyping panels were applied in the SA Drakensberger, a high proportion of low minor allele frequency (MAF) SNP as well as weak LD between SNP was observed (Qwabe et al., 2013; Zwane et al., 2016; Lashmar et al., 2018). It is therefore expected that the lower limit of SNP required on a reduced density panel will be higher than proposed for taurine breeds in which longer haplo-blocks of these SNP exist. Therefore, while approaches to SNP selection for reduced genotype panels has been presented for purebred populations (Judge et al., 2016), these strategies may not be applicable in admixed populations where, in particular, LD characteristics may differ. Careful consideration of the genomic characteristics of individual SNP, in terms of the selection strategy, is necessary to generate a reduced panel with optimal utility in admixed populations. The strategies considered within this study to develop lower density panels may also be useful in informing such strategies in other breeds with no relationship to Sanga-type animals.

Various methods have been proposed to identify the most appropriate low-density SNP sets from a larger pool of existing SNP using certain genomic characteristics as inclusion criteria. These methods have generally considered genomic parameters such as mean MAF (Corbin et al., 2014; Judge et al., 2016), and inter-marker relationship estimates, i.e., LD (Ogawa et al., 2016), while maintaining more or less an even dispersion of selected SNP across the genome. The most efficient selection strategy will ensure a minimization of the size of the reduced SNP panel without compromising imputation accuracy and hence facilitate the use of such a panel in imputation-driven applications for local breeds. Even though the cost structure of genotyping with SNP panels depends on many different factors, minimizing the number of SNP to a low enough density may enable (1) a transition to alternative technologies in the future (e.g., genotype-by-sequencing) or (2) including multiple species on a single platform which would contribute to a larger purchase order and thus a reduction in price per panel.

The overarching aim of the present study was to determine the accuracy of imputing from low- to high-density panels in the indigenous SA Drakensberger breed. The main objectives were achieved by varying (1) the SNP density and (2) the selection strategy of SNP for various custom-derived low-density panels. An additional objective was to determine the impact of relatedness between the validation and reference populations on the achievable imputation accuracy.

Materials and Methods

Ethical approval to perform this study was obtained from the University of Pretoria's Faculty of Natural and Agricultural Science (EC151106-024). The SNP genotypes used were generated as part of the Beef Genomics Project (BGP; http://www.livestockgenomics.co.za/) and written consent was given by the Drakensberger Breeders' Society to use the SA Drakensberger genomic and phenotypic data.

Pedigree and genotype data

Pedigree and SNP data were available for 1,135 SA Drakensberger cattle originating from 48 breeders and born between 1982 and 2017. Available pedigree information for the breed was obtained from SA Stud Book (SA Stud Book Association) and consisted of 6,074 animals, which comprised the genotyped animals and their ancestors. Genotypes generated using the GeneSeek Genomic Profiler Bovine 150K SNP panel (Neogen Corporation, Lansing, MI) consisting of 139,480 SNP were available for 214 male and 921 female SA Drakensberger cattle, all of which had a sample call rate exceeding 90%. SNP were mapped to the UMD3.1 bovine reference genome using SNPchiMp versions 3 and SNPConvert software (Nicolazzi et al., 2016). Only SNP mapped to autosomes (*Bos Taurus* autosome; BTA1–BTA29) and with call rates exceeding 95%, MAF exceeding 1% and that did not violate Hardy Weinberg Equilibrium ($P < 0.01 \times 10^{-6}$) were retained for downstream analysis. A total of 120,608 SNP (mean call rate = 99.4%) remained after edits.

Parent mismatches within the genotype data set were identified as pedigree-defined putative parent–progeny pairs displaying >2% Mendelian inconsistencies. Where parentage errors were detected, the parents were set to missing in the pedigree file if no alternative match within the genotype data set was identified; the pedigree was updated if an alternative match was detected. The data set was then separated into a reference population (n = 900) and a validation population (n = 235). The validation population consisted of the youngest animals but with no more than 3 paternal sibs represented. The reference population was used to estimate SNP MAF and the extent of LD, both of which were used as criteria for the selection of SNP in the low-density panels (discussed hereafter), and to model haplotypes used in imputation.

SNP selection methods

Different strategies were used to develop several custom-derived low-density SNP panels consisting of 2,500, 5,000, 10,000, 20,000, and 50,000 SNP. The number of SNP selected per autosome was proportional to the length of each autosome; therefore, more SNP were selected for longer autosomes. The number of SNP selected per autosome to fulfill each of the different panel densities is summarized in Table 1. Six alternative algorithms were used to generate the custom SNP panels, and these were implemented as follows.

Random selection (RAN)

The predefined number of SNP required per autosome was chosen at random until each of the respective panel densities was reached.

Mid-point, equidistant selection (MID)

The length of each autosome, defined as the difference in base pairs between the first and last genotyped SNP per autosome, was divided into equally sized segments and the SNP closest to the physical midpoint of each segment was chosen. The segment size per autosome was calculated as the autosomal length divided by $n-1$, where n is the predefined number of SNP to be chosen for that specific autosome. Due to an uneven distribution of

Table 1. The number of SNP that were selected per chromosome for the 2,500, 5,000, 10,000, 20,000, and 50,000-marker low-density genotyping panels

		Number of SNPs					
Chromosome	Chromosome length (Mb)	HD[1]	2,500	5,000	10,000	20,000	50,000
1	158.21	7,371	152	305	610	1,222	3,166
2	136.62	6,455	133	267	535	1,070	2,734
3	121.38	5,796	120	240	481	961	2,428
4	120.55	5,582	115	231	463	925	2,414
5	121.14	6,041	125	250	501	1,001	2,417
6	119.40	6,626	137	274	549	1,098	2,385
7	112.60	5,705	118	236	473	946	2,230
8	113.35	5,157	106	213	428	855	2,254
9	105.59	4,948	102	205	410	820	2,096
10	104.23	4,891	101	202	406	811	2,075
11	107.24	5,028	104	208	417	833	2,137
12	91.10	4,211	87	174	349	698	1,797
13	84.20	3,924	81	162	325	650	1,678
14	84.03	4,701	97	194	390	779	1,679
15	85.20	3,984	82	165	330	660	1,700
16	81.65	3,741	78	156	310	621	1,623
17	75.11	3,446	72	143	286	572	1,491
18	65.87	3,073	64	128	255	510	1,313
19	63.89	2,935	61	122	243	487	1,269
20	71.88	3,799	78	158	315	630	1,440
21	71.53	3,338	70	139	277	554	1,416
22	61.24	2,901	61	121	241	482	1,226
23	52.45	2,469	52	103	205	410	1,045
24	62.59	3,377	70	140	280	560	1,254
25	42.76	2,003	42	84	166	333	861
26	51.58	2,469	52	103	205	410	1,029
27	45.35	2,127	45	89	176	353	899
28	46.24	2,131	45	89	177	354	922
29	51.17	2,379	50	99	197	395	1,022

[1]HD = 120,608 SNP that remained after quality control procedures.

SNP in specific autosomal regions after quality control, certain segments did not contain any candidate SNP and, in these situations, the SNP on the boundary closest to the starting position of the adjacent segment was chosen.

Equidistant selection that maximized MAF (DISTMAF)

SNP within segments of equal size were chosen based on an index that maximized MAF whilst attempting to adhere to the ideal inter-SNP spacing per autosome for each panel density. The SNP with the highest MAF was chosen within the first segment per autosome after which each SNP within subsequent segments was chosen based on the highest index score calculated as proposed by Matukumalli et al. (2009) and Zhang and Druet (2010):

$$score_i = MAF_{SNP_i} \left[ssize - \left|ssize - d_{SNP_i, SNP_{i-1}}\right|\right]$$

where $\mathrm{MAF}_{\mathrm{SNP}_i}$ represented the MAF of a candidate SNP *i*, *s*size represented the genomic length of each segment within a given autosome and represent $d_{\mathrm{SNP}_i,\,\mathrm{SNP}_{i-1}}$ represented the genomic distance between the base pair position of a candidate $\mathrm{SNP}_i \text{ and } \mathrm{SNP}_{i-1}$, where SNP_{i-1} is the SNP selected in the previous segment. If a segment contained no SNP, a second SNP was chosen in the next segment, i.e., the SNP with the second highest index score was also chosen.

Segment-based selection combining MAF and LD (MAFLD)

An index score combining MAF and LD information was calculated per SNP within genomic segment. The MAF and LD per SNP were first standardized so that the weights on each attribute within segment were equal before summation. The scores were derived as follows:

$$\mathrm{score}_{ij} = \frac{\mathrm{MAF}_{\mathrm{SNP}_i}}{\mathrm{SD}_{\mathrm{MAF}_{\mathrm{seg}_j}}} + \frac{\overline{\mathrm{LD}_{\mathrm{SNP}_{ij}}}}{\mathrm{SD}_{\mathrm{LD}_{\mathrm{seg}_j}}}$$

where $\mathrm{MAF}_{\mathrm{SNP}_i}$ represented the MAF of candidate SNP *i* in segment $j, \mathrm{SD}_{\mathrm{MAF}_{\mathrm{seg}_j}}$ represented the standard deviation for MAF in segment $j, \overline{\mathrm{LD}_{\mathrm{SNP}_{ij}}}$ represented the mean LD between candidate SNP_i and all other SNP within segment *j*, and $\mathrm{SD}_{\mathrm{LD}_{\mathrm{seg}_j}}$ represented the standard deviation for all pairwise LD interactions within segment *j* . Within each segment, the SNP with the highest index score was chosen. A second SNP was selected in the segments at both ends of each autosome and the number of segments was therefore equal to the number of SNP to be chosen −2. The second SNP in the peripheral chromosomal segments was selected based on a score combining MAF and the partial correlation of that SNP with all remaining candidate SNP in the periphery segment. Adjustments were made to the partial correlation to account for the relationship between each candidate SNP and the SNP already selected in the initial round of selection. This calculation was performed according to Judge et al. (2016) as follows:

$$r(SNP_i, SNP_j|SNP_{sel}) = \frac{r(SNP_i, SNP_j) - r(SNP_i, SNP_{sel})\, r(SNP_j, SNP_{sel})}{[1 - r^2(SNP_i, SNP_{sel})]^{1/2}[1 - r^2(SNP_j, SNP_{sel})]^{1/2}}$$

where $r(\mathrm{SNP}_i, \mathrm{SNP}_j \mid \mathrm{SNP}_{\mathrm{sel}})$ represented the partial correlation between candidate SNP_i and candidate SNP_j corrected for the correlation with the already selected SNP, SNP_{sel}. The terms $(\mathrm{SNP}_i, \mathrm{SNP}_j), r(\mathrm{SNP}_i, \mathrm{SNP}_{\mathrm{sel}}), \text{and } r(\mathrm{SNP}_j, \mathrm{SNP}_{\mathrm{sel}})$ represented the correlations between 2 candidate SNP, between the first candidate SNP and the selected SNP, and between the second candidate SNP and the selected SNP, respectively.

Partitioning-around-medoids (PAM)

SNP on each autosome were partitioned into a number of clusters based on their proximity in genomic position using the PAM algorithm implemented in R's "*clara*" package (Kaufman & Rousseeuw, 2009). The number of clusters was equal to the number of pre-defined SNP to be selected per autosome. The medial SNP within each SNP cluster was chosen.

Intra-autosomal LD-based clustering algorithm (LDCLUST)

SNP were selected within each autosome separately using a hierarchical clustering algorithm that clusters subsets of SNP according to a prespecified LD threshold between SNP. The SNP density of a given panel is therefore adjusted according to local LD levels and is not preselected to achieve even genomic dispersion. This algorithm was executed using the SS4I software (Hérault et al., 2016) as implemented by Herry et al. (2018) for layer chickens. The SNP density of these panels in the present study was therefore not pre-defined but corresponded to the number of LD-based clusters identified, the number of which was determined by the different r^2 thresholds imposed (i.e., r^2 = 0.05, 0.1, and 0.2). The SNP with the greatest MAF within each of the defined clusters using this method was subsequently selected to develop the panels in the present study. The choice of LD thresholds was based on weak autosome-level LD previously reported for the SA Drakensberger breed (mean pairwise r^2 ranged from 0.11 for BTA28 to 0.17 for BTA14 between SNP pairs separated by ≤100kb using a 120,218 SNP panel with a mean inter-SNP distance of 20.9kb; Lashmar et al., 2018).

Imputation and imputation accuracy

Imputation from each of the low-density panels to the higher density was performed using FImpute version 2.2 software (Sargolzaei et al., 2014) based on both pedigree information and population-wide LD. Imputation with this software is carried out simultaneously on a per chromosome basis. The software's default settings were used with regards to specifications of the sliding window used to capture haplotype similarities (i.e., shrink factor = 0.150 and overlap = 0.650).

Imputation accuracy was quantified using 3 parameters namely: (1) genotype concordance rate (GCR), (2) allele concordance rate (ACR), and (3) the Pearson correlation between the true and imputed genotypes (COR). These parameters were averaged per animal (ACR_{ANIM},

GCR_{ANIM}, and COR_{ANIM}, respectively) and per SNP (ACR_{SNP}, GCR_{SNP}, and COR_{SNP}, respectively). The GCRs and ACRs were calculated as the proportion of correctly imputed genotypes and alleles, respectively. For genotype concordance, a score of zero was given to a SNP if it had either 1 allele (homozygous true vs. heterozygous imputed) or both alleles (opposing homozygous for true vs. imputed) incorrectly imputed. For allele concordance, a score of 0.5 was given if 1 allele was correctly imputed, i.e., a homozygous true genotype vs. a heterozygous imputed genotype. For both these measures, concordance was calculated (1) for only masked genotypes and (2) for both masked and unmasked genotypes, i.e., all of the genotypes in the high-density panel. Unmasked genotypes were the genotypes that corresponded to each of the in silico selected SNP panels while masked genotypes were the genotypes that corresponded to the higher density panel and were intentionally excluded from each of the in silico SNP panels. The latter calculation was included to mimic what would be expected to happen in real life (Judge et al., 2016).

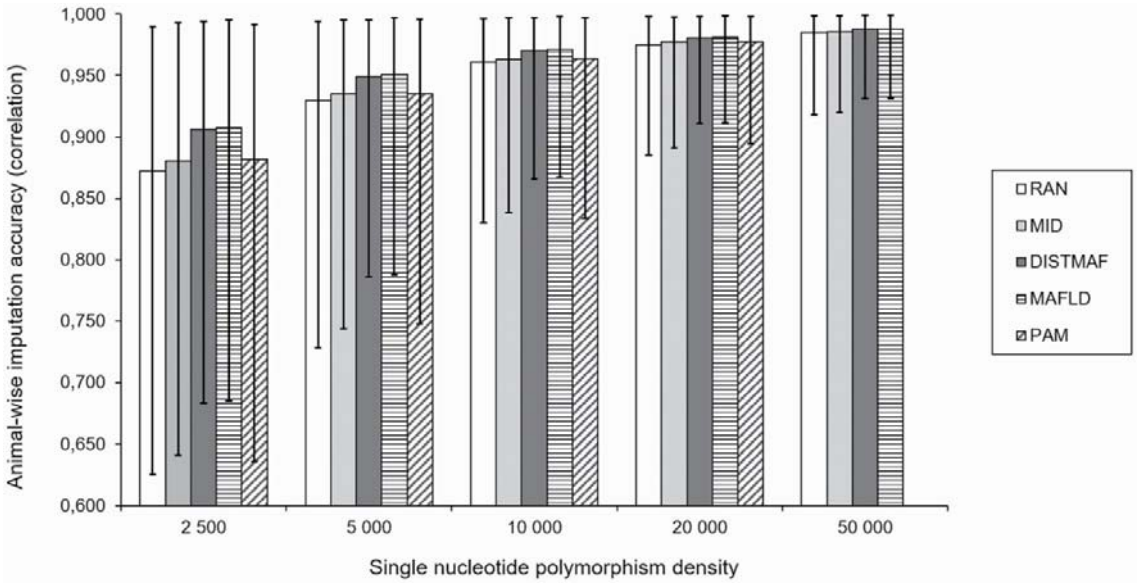

Figure 1. Mean correlation-based imputation accuracies (COR_{ANIM}) for different genotyping panel densities derived using 5 different SNP selection methodologies (RAN: random selection, MID: midpoint selection, DISTMAF: equidistant selection maximizing MAF, MAFLD: combinative selection for MAF and LD, PAM: partitioning-around-medoids selection). Error bars represent minimum and maximum COR_{ANIM}.

Results

Number of SNP on the lower density panel

The mean imputation accuracy per animal increased with increasing panel density but did so at a diminishing rate (Figure 1). Animal-wise correlation-based imputation accuracy, COR_{ANIM}, ranged (minimum to maximum) from 0.625 to 0.990, 0.728 to 0.994, 0.830 to 0.996, 0.885 to 0.998, and 0.918 to 0.999 when 2,500, 5,000, 10,000, 20,000, and 50,000 SNP were randomly chosen. Mean COR_{ANIM} increased by 0.055, 0.043, and 0.043 units for the MID, DISTMAF, and MAFLD methods, respectively, when the number of SNP doubled

from 2,500 to 5,000 SNP. Mean COR_{ANIM} increased by only 0.008, 0.007, and 0.007 for MID, DISTMAF, and MAFLD, respectively, when the density increased from 20,000 to 50,000 SNP. Moreover, the standard deviation for COR_{ANIM} reduced with increasing panel density; the standard deviation reduced from 0.075 for COR_{ANIM} to 0.014 as panel density increased from 2,500 (mean COR_{ANIM} = 0.872) to 50,000 SNP (mean COR_{ANIM} = 0.985). The difference between ACR_{ANIM} and GCR_{ANIM} was, for example, 0.051 units for the 2,500 SNP panel vs. 0.007 units for the 50,000 SNP panel when the DISTMAF SNP selection strategy was used.

SNP selection method

When the LDCLUST algorithm was used, SNP panels consisting of 23,469, 30,681, and 41,403 SNP were generated corresponding to the LD thresholds of r^2 = 0.05, r^2 = 0.1 and r^2 = 0.2, respectively. The mean COR_{ANIM} (standard deviation), ACR_{ANIM} (standard deviation), and GCR_{ANIM} (standard deviation) based on the 23,469 SNP panel, which was somewhat comparable in density to the predefined 20,000 SNP panels of the other methods, was 0.979 (0.018), 0.987 (0.11), and 0.975 (0.021), respectively. These values compared favorably with both the DISTMAF and LDMAF methods (e.g., ACR_{MASKED} = 0.987 vs. 0.988 for both DISTMAF and LDMAF; Table 2), and were superior to the accuracy estimates for all the remaining methods. Furthermore, the mean COR_{ANIM} (standard deviation), ACR_{ANIM} (standard deviation), and GCR_{ANIM} (standard deviation) values were 0.982 (0.016), 0.989 (0.010), and 0.979 (0.019) when the 30,681 SNP panel was used and improved to 0.986 (0.013), 0.991 (0.008), and 0.982 (0.016) when the 41,403 SNP panel was used. Because the SNP density of the panels developed using the LDCLUST method differed from the methods that used a predefined number of SNP, all subsequent results relate only to the SNP selection methods that had a predefined

Across all panel densities evaluated, the poorest imputation accuracy was always achieved when the SNP were randomly selected. Strategies that were based on the selection of SNP using scores combining MAF with other attributes (i.e., DISTMAF and MAFLD) outperformed the other selection strategies; the MAFLD method always resulted in the best imputation accuracy (Table 2) irrespective of panel density.

Negligible differences in imputation accuracy existed between the SNP selection strategies that only considered the genomic location of SNP (i.e., MID and PAM). The DISTMAF and MAFLD methods were the only selection strategies with mean imputation accuracies exceeding 0.970 at a density of 5,000 SNP; ACR_{ANIM} ranged from 0.856 to 0.997 and from 0.859 to 0.997 for DISTMAF and MAFLD, respectively, at 5,000 SNP. The improvements in imputation accuracy were marginal when unmasked SNP were included in the calculation of GCR_{ANIM} and ACR_{ANIM}, increasing by a mean of 0.004 (SD = 0.002) and 0.002 (SD = 0.001), respectively, across all densities and SNP selection strategies.

Degree of relatedness between validation and reference populations

The genomic relationship between a given animal in the validation population with animals in the reference population directly impacted the imputation accuracy. A scatter plot of each validation animal's imputation accuracy and the mean of its top 10 coefficients of relatedness to the reference population are shown in Figure 2.

Table 2. GCRs and ACRs for different low-density SNP panels

		Imputation accuracy			
LD panel	Method	GCR_{MASKED} (SD)[1]	GCR_{ALL} (SD)[2]	ACR_{MASKED} (SD)[1]	ACR_{ALL} (SD)[2]
2,500	RAN	0.857 (0.075)	0.860 (0.074)	0.925 (0.041)	0.926 (0.040)
	MID	0.865 (0.075)	0.868 (0.073)	0.929 (0.041)	0.931 (0.040)
	DISTMAF	0.893 (0.066)	0.896 (0.065)	0.944 (0.036)	0.945 (0.035)
	LDMAF	0.895 (0.066)	0.897 (0.065)	0.945 (0.035)	0.946 (0.035)
	PAM	0.867 (0.074)	0.869 (0.072)	0.930 (0.040)	0.931 (0.040)
5,000	RAN	0.918 (0.052)	0.922 (0.050)	0.958 (0.028)	0.959 (0.027)
	MID	0.922 (0.066)	0.925 (0.064)	0.959 (0.042)	0.960 (0.041)
	DISTMAF	0.940 (0.044)	0.943 (0.042)	0.969 (0.023)	0.970 (0.022)
	LDMAF	0.942 (0.042)	0.944 (0.022)	0.970 (0.022)	0.971 (0.021)
	PAM	0.925 (0.050)	0.928 (0.048)	0.961 (0.027)	0.963 (0.026)
10,000	RAN	0.954 (0.035)	0.957 (0.032)	0.976 (0.018)	0.978 (0.017)
	MID	0.954 (0.055)	0.957 (0.054)	0.975 (0.037)	0.977 (0.037)
	DISTMAF	0.964 (0.029)	0.967 (0.026)	0.982 (0.015)	0.983 (0.014)
	LDMAF	0.966 (0.028)	0.968 (0.025)	0.982 (0.014)	0.984 (0.013)
	PAM	0.957 (0.034)	0.960 (0.031)	0.978 (0.017)	0.980 (0.016)
20,000	RAN	0.970 (0.025)	0.975 (0.021)	0.985 (0.013)	0.987 (0.011)
	MID	0.972 (0.023)	0.977 (0.019)	0.986 (0.012)	0.988 (0.010)
	DISTMAF	0.976 (0.020)	0.980 (0.017)	0.988 (0.010)	0.990 (0.009)
	LDMAF	0.977 (0.020)	0.980 (0.017)	0.988 (0.010)	0.990 (0.009)
	PAM	0.973 (0.023)	0.977 (0.019)	0.986 (0.012)	0.988 (0.010)
	LDCLUST[3]	0.975 (0.021)	0.980 (0.017)	0.987 (0.110)	0.990 (0.009)
50,000	RAN	0.981 (0.016)	0.989 (0.010)	0.991 (0.008)	0.994 (0.005)
	MID	0.982 (0.016)	0.990 (0.009)	0.991 (0.008)	0.995 (0.005)
	DISTMAF	0.985 (0.014)	0.991 (0.008)	0.992 (0.007)	0.995 (0.004)
	LDMAF	0.985 (0.014)	0.991 (0.008)	0.992 (0.007)	0.995 (0.004)
	PAM	0.982 (0.015)	0.990 (0.009)	0.991 (0.008)	0.995 (0.005)

[1]GCR_{ALL} and ACR_{ALL} = mean imputation accuracy across the full set of 120,608 SNP, including both true and imputed SNP.

[2]GCR_{MASKED} and ACR_{MASKED} = mean imputation accuracy across masked SNP only, i.e., only the SNP imputed per density.

[3]Exact SNP panel density was 23,469 SNP.

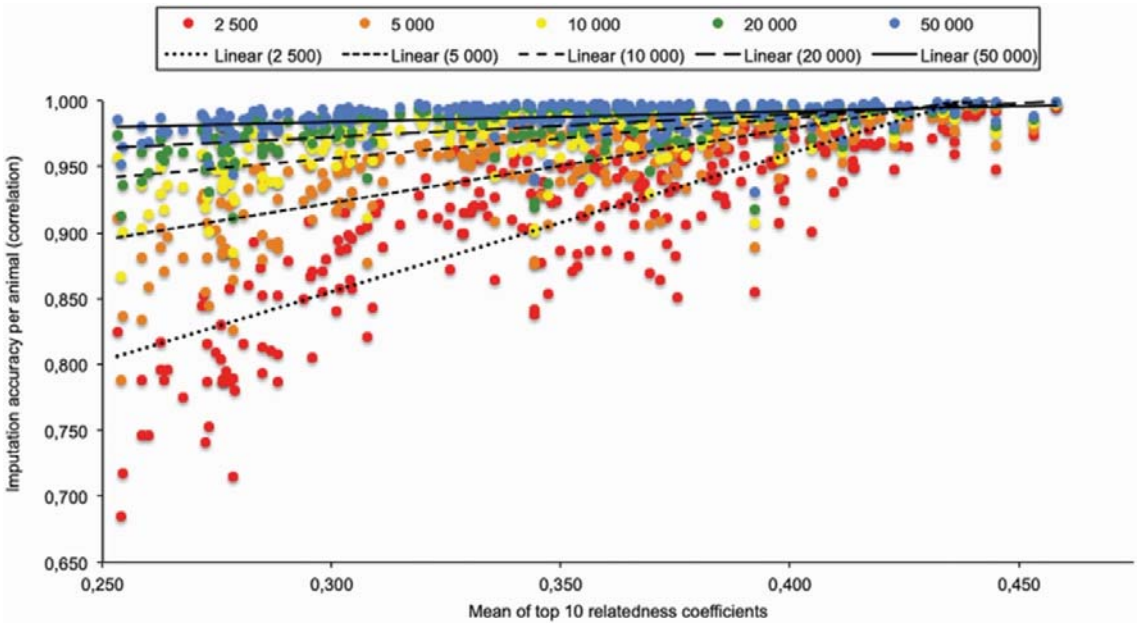

Figure 2. The validation animal-wise relationship between relatedness to the reference population (mean of top 10 coefficients) and imputation accuracy from different panel densities. The linear regression lines for imputation accuracy on relatedness are illustrated for each panel density.

The influence of genomic relatedness between reference and validation animals on the correlation-based accuracy was more pronounced when fewer SNP were included on the lower density panel. A minimum (maximum) COR_{ANIM} of 0.825 (0.995), 0.911 (0.997), 0.957 (0.998), 0.974 (0.998), and 0.986 (0.999) was observed for animals that were least (most) related, respectively, to the reference population when 2,500, 5,000, 10,000, 20,000, and 50,000 SNP were utilized. The regression coefficient (simple linear regression) of mean animal-wise imputation accuracy (correlation) on mean relatedness to the reference population (top 10 relatives), reduced as panel density increased (data not shown). The corresponding R^2 value for the 2,500, 5,000, 10,000, 20,000, and 50,000 SNP panels was 0.684, 0.557, 0.406, 0.265, and 0.135, respectively. Therefore, a larger proportion of the variability in COR_{ANIM} was explained by animal relatedness to the reference population when a lower density panel was used.

Imputation accuracy per SNP

Variation between autosomes

Imputation accuracy, depicted by either mean genotype- (GCR_{SNP}) or allele concordance rates (ACR_{SNP}), differed by autosome (Figure 3). Using the 10,000 SNP panel as an example, the lowest mean ± SD ACR across masked SNP was for BTA26 (0.971 ± 0.021) when random selection was undertaken. When other SNP selection strategies were employed, imputation accuracy of BTA19 (MID = 0.971 ± 0.018; DISTMAF = 0.976 ± 0.015) and BTA23 (MAFLD = 0.978 ± 0.015; PAM = 0.971 ± 0.020) were the worst of all chromosomes. The greatest chromosome-wide allele concordance was for BTA5 for all selection strategies except the MAFLD strategy; for the MAFLD strategy, BTA24 (0.985 ± 0.12) was the best.

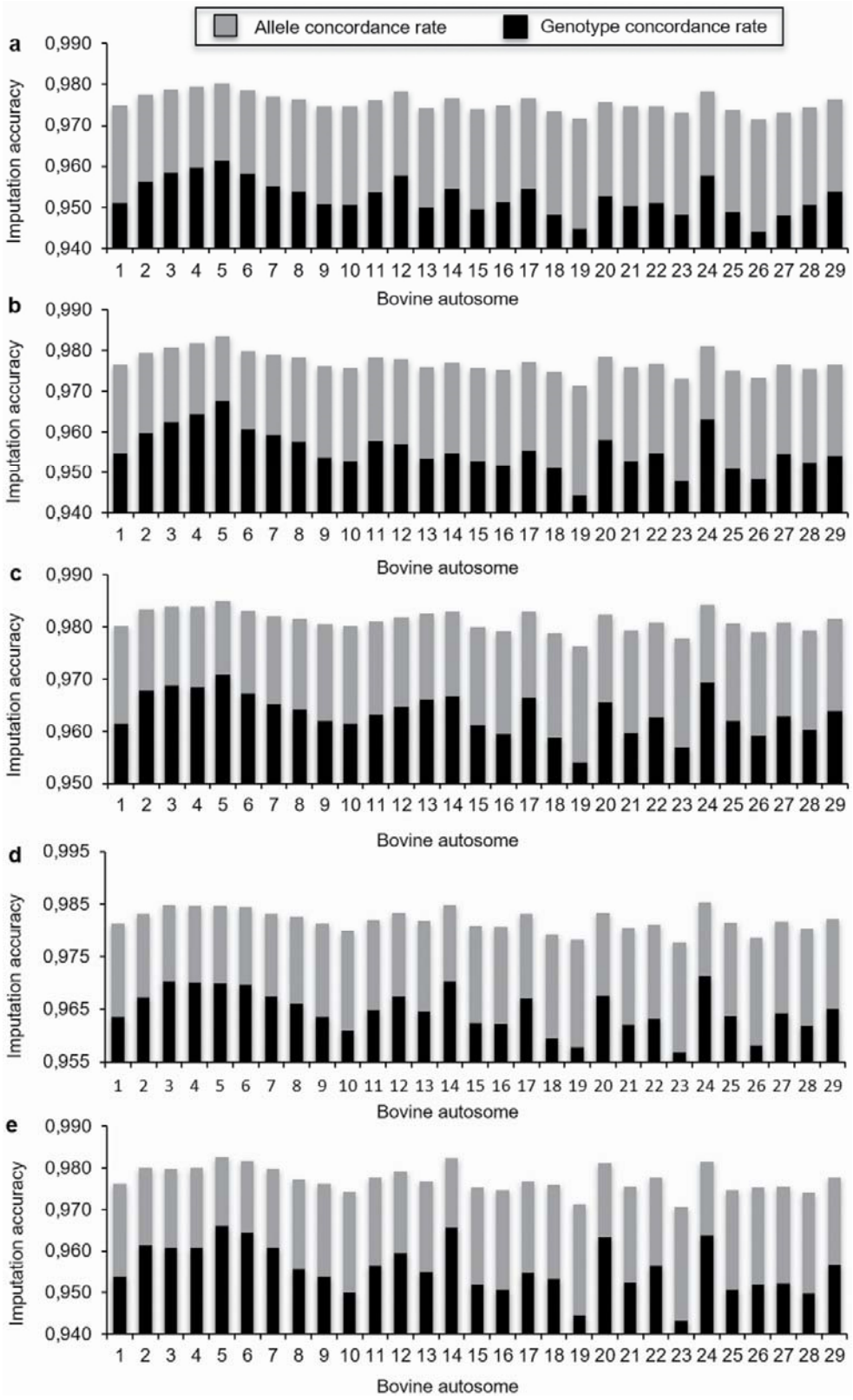

Figure 3. Mean concordance-based imputation accuracy measures for 10,000 SNP panels derived using the 5 different marker selection methodologies (a: random selection, b: midpoint selection, c: equidistant selection maximizing MAF, d: combinative selection for MAF and LD, e: partitioning-around-medoids selection).

Variation within autosomes

Within autosomes, variability in SNP-level imputation accuracy existed by location relative to the center of the autosome. The distribution of COR_{SNP} across individual autosomes is illustrated in Figure 4 for the 2,500 and 50,000 SNP panels where the SNP were randomly selected. The pattern of COR_{SNP} distribution on many autosomes indicated a tendency towards poorer imputation accuracy at the autosomal peripheries. For the 2,500 and 50,000 SNP panels, mean ± SD COR_{SNP} for SNP located in the central 1 Mb of autosomes (0.5 Mb each side of the physical midpoint of each autosome) were 0.781 ± 0.067 and 0.968 ± 0.015 across all autosomes, respectively, when RAN was used. The corresponding values for the two 1 Mb autosomal peripheries were 0.682 ± 0.149 and 0.664 ± 0.07, and 0.938 ± 0.130 and 0.968 ± 0.018, for the 2,500 and 50,000 SNP panels, respectively. Using a 10,000 SNP panel, the mean ± SD COR_{SNP} for SNP located within 1 Mb of the start, the center and the end of autosomes was 0.923 ± 0.036, 0.943 ± 0.031, and 0.912 ± 0.030, respectively, across all autosomes. For MAFLD, using 10,000 SNP, the mean COR_{SNP} for the central and peripheral regions (1 Mb) per autosome are depicted in Table 3.

Table 3. Mean COR_{SNP} per autosome for SNP located on the autosomal extremities and within the center of the autosomes when selection combining minor allele frequency and LD (MAFLD) was used to derive a 10,000 SNP panel

Autosome	First	Centre	Last
1	0.875	0.942	0.918
2	0.945	0.950	0.910
3	0.916	0.978	0.914
4	0.866	0.943	0.912
5	0.953	0.945	0.923
6	0.963	0.954	0.919
7	0.845	0.955	0.930
8	0.938	0.970	0.888
9	0.986	0.937	0.927
10	0.910	0.941	0.945
11	0.930	0.961	0.802
12	0.926	0.938	0.911
13	0.939	0.944	0.929
14	-[1]	0.945	0.904
15	0.881	0.926	0.951
16	0.914	0.976	0.894
17	0.947	0.959	0.933
18	0.879	0.916	0.858
19	0.952	0.941	0.915
20	0.935	0.968	0.934
21	0.874	0.966	0.898
22	0.925	0.939	0.936
23	0.983	0.970	0.911
24	0.885	0.955	0.935
25	0.955	0.924	0.911
26	0.908	0.809	0.919
27	0.921	0.912	0.932
28	0.934	0.931	0.929
29	0.956	0.958	0.858

Autosomal extremities were defined as first and last 1 Mb of each autosome, whereas the autosomal center was defined as 0.5 Mb to either side of the physical midpoint of each autosome.

[1]No SNP were mapped to the first 1 Mb of BTA14 after quality control.

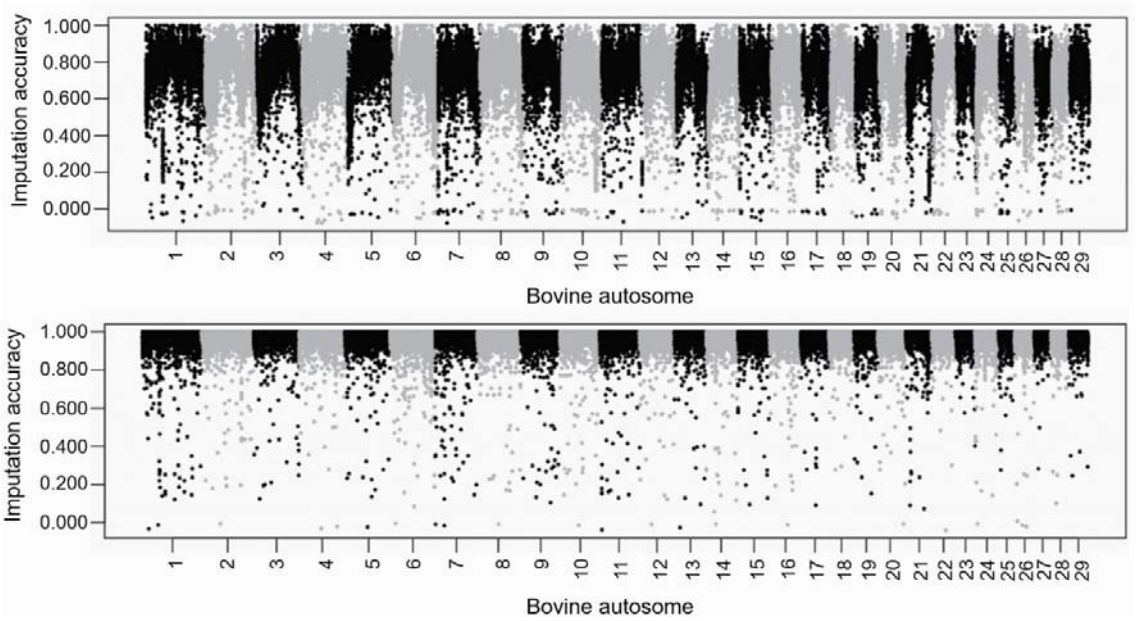

Figure 4. Marker-wise imputation accuracy (correlation) per autosome for 2,500- (top) vs. 50,000 (bottom) SNP panels derived by random selection.

Variability in imputation accuracy based on SNP MAF

Imputation accuracy differed by SNP MAF. The nature of the relationship between MAF and imputation accuracy, however, varied depending on the accuracy statistic used (Figure 5). Mean values of concordance measures reduced whilst the correlation measure strengthened with increasing MAF. The mean COR_{SNP} for SNP classified in the highest MAF bin ($0.4 < MAF \leq 0.5$) was 0.120 and 0.135 units higher than for SNP classified in the lowest MAF bins ($0.01 < MAF \leq 0.02$) when the RAN and MAFLD selection strategies were respectively used. For the concordance measures, the mean imputation error rate for GCR_{SNP}, defined as one minus mean GCR_{SNP}, was approximately double the error rate for mean ACR_{SNP}, defined as one minus mean ACR_{SNP}, for all MAF bins. The difference between the allele- (AER) and genotype error rates (GER) $\Delta_{\mathrm{AER_{SNP}},\,\mathrm{GER_{SNP}}}$, per SNP was calculated as follows according to Ma et al. (2013):

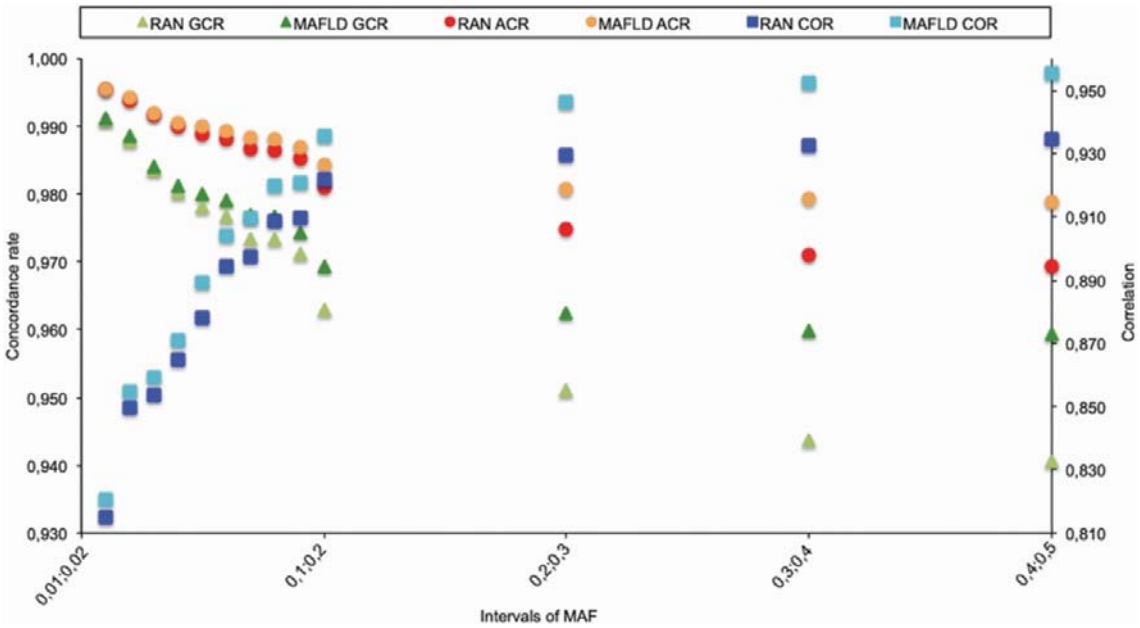

Figure 5. Mean genotype and ACR (primary Y-axis) as well as correlation-based (secondary Y-axis) imputation accuracy for intervals of increasing MAF. Points on the graph below 0.1 MAF represent mean imputation accuracy for MAF intervals increasing in increments of 0.01. (RAN: random selection, MAFLD: combinative selection for MAF and LD).

$$\Delta_{\mathrm{AER,GER}} = 2(1 - \mathrm{ACR}) - (1 - \mathrm{GCR})$$

The difference increased from zero for the lowest MAF bin (0.01 < MAF ≤ 0.02) to 0.002 (RAN) and 0.001 units (MAFLD) for the highest MAF bin (0.4 < MAF ≤ 0.5). The discrepancy between both concordance-based measures and the correlation-based accuracy measure was more pronounced when MAF was low but diminished as MAF increased. Monomorphic SNP and SNP with MAF < 0.01 were removed during quality control and were therefore not considered.

Discussion

The sustainability of routine genotyping, where the breeder incurs the full cost, will necessitate a transition to a low-density SNP panel that contains fewer SNP, especially in countries with poorer financial resources available for the commercial deployment of genomic technologies. The reduced associated cost of lower density panels may, therefore, improve the uptake of such technologies by more beef farmers. The principal motivation of this study was therefore to determine the achievable imputation accuracy from various custom-derived lower-density SNP panels for a beef cattle breed characterized by genomic admixture, as is the case for a large proportion of African cattle. The investigation in an admixed population is the novelty in the present study; similar exercises have been untaken in purebred populations but LD patterns differ across breeds and the rate of decay in LD is faster in admixed populations (Shifman, 2003; Toosi et al., 2010). While this difference in LD patterns can impact SNP selection, it will also impact the ability of a reduced SNP panel to fully capture all segregating haplotypes thus impacting the downstream imputation accuracy.

Genotyping panel density

The trend of increasing imputation accuracy in the more densely populated SNP panels in the admixed Drakensberger population is consistent with previous studies in non-admixed cattle (Zhang and Druet, 2010; Mulder et al., 2012; Carvalheiro et al., 2014; Judge et al., 2016) and sheep (Hayes et al., 2012; O'Brien et al., 2019), and is likely attributable to the fact that haplotypes were more easily and accurately resolved when a greater number of unmasked, neighboring SNP were available or, in other words, there were fewer SNP to impute (Tsai et al., 2017). The ever diminishing improvement in imputation accuracy with each incremental increase in panel density size is consistent with previous results (Judge et al., 2016; O'Brien et al., 2019), suggesting that at higher SNP densities (>10,000 SNP panel), the density was already adequate to resolve shared haplotypes and to achieve high imputation accuracy even in highly admixed populations as is the case for the Drakensberger. Increasing the panel density to 20,000 SNP or more, had a negligible benefit on imputation accuracy and is expected to be less cost effective for less popular breeds in developing countries where the service costs per genotyping assay are still considered high relative to more developed countries.

Across all 5 selection methods that were based on a predefined number of SNP, the average improvement in COR_{ANIM} was 0.05, 0.03, and 0.01 units when panel densities were doubled from 2,500 to 5,000, 5,000 to 10,000, and 10,000 to 20,000 SNP, respectively. This was in agreement with results reported by Judge et al. (2016) in Irish cattle who documented improvements in animal-wise correlations of 0.07, 0.02, and 0.01 units when the number of markers was doubled from 1,000 to 2,000, 3,000 to 6,000, and 6,000 to 12,000 SNP, respectively. Carvalheiro et al. (2014) documented similarly small gains in correlation-based accuracy (0.01 units) of imputation to 777,000 SNP when the density of their custom SNP panels was increased from 11,000 to 48,000 SNP in Nellore beef cattle.

Relative to the results of the present study, Aliloo et al. (2018) reported poorer imputation accuracy for East African crossbred dairy cattle which have weak LD levels similar to that of the SA Drakensberger (Lashmar et al., 2018). At SNP densities of 4,725, 11,773, and 19,812, which were comparable to our 5,000, 10,000, and 20,000 SNP panels, Aliloo et al. (2018) documented GCRANIM values of 0.822, 0.896, and 0.923 (compared with 0.940, 0.964, and 0.976 for DISTMAF in our study) when SNP were selected based on maximizing MAF within equal sized genomic segments. The lower imputation accuracy in the study of Aliloo et al. (2018) may be attributable to the fact that imputation was calculated (1) across many admixed breeds with small sample sizes per breed and (2) based on imputation to a much higher density (777,000 SNP). Makina et al. (2015) reported a weak sharing of long-distance LD among Sanga breeds; therefore imputation accuracy across Sanga breeds is expected to be lower and more comparable to the accuracies reported by Aliloo et al. (2018).

For every unit increase in mean relatedness to the top 10 closest relatives in the reference population, the unit increase in animal-wise imputation accuracy became progressively smaller as panel density increased (e.g., minimum correlation coefficient = 0.081 at 50,000 SNP). This trend suggests that the impact of across-population relatedness on imputation accuracy reduced as the number of SNP on the genotype panel became sufficiently dense. Imputation accuracy in the more dense SNP panels increasingly became a function of shared

population-wide LD between SNP rather than pedigree-defined relationships. This trend was in agreement with previous studies in cattle (Ventura et al., 2014; García-Ruiz et al., 2015) including those on admixed populations (Júnior et al., 2017).

SNP selection strategies

The SNP selection strategies that considered MAF and an additional attribute, either inter-SNP distance or LD, generally produced more accurate imputation than strategies that chose SNP either randomly or based solely on genome-wide SNP distribution. Moreover, the difference in mean imputation accuracy between methods was more pronounced at lower panel densities consistent with reported elsewhere in purebred populations (Judge et al., 2016). Judge et al. (2016) documented 0.031 vs. 0.004 units higher accuracy for "block" selection, which was carried out the same as MAFLD in the present study, as opposed to random selection; these selections methods were also the best and worst methods in the Judge et al. (2016) study. A major contributing factor to the improvement in imputation accuracy in the present study was the enrichment of chromosomal extremities when MAFLD was used; while a second SNP was selected in the first and last chromosomal segments, no SNP were selected in these regions when RAN was used. Selection using MAFLD was furthermore undertaken within genomic segments of even size, thereby securing an even distribution across the entirety of autosomes. Considerably more variability in the distance between adjacent SNP was detected among SNP selected using the RAN method. For the 10,000 SNP panels, the mean ± SD distance between adjacent SNP, in kilobase pairs (kb), was 250.30 ± 266.03, 251.12 ± 35.85, 251.00 ± 100.05, 251.10 ± 49.95, and 251.16 ± 101.69 units when the RAN, MID, PAM, DISTMAF, and MAFLD methods were employed.

SNP selection methods combining both attributes of MAF and LD have previously been reported to generate the most accurate imputation in cattle compared with other methods such as random selection, selection using distance scores, or selection using machine-learning algorithms (Carvalheiro et al., 2014; Judge et al., 2016). Other studies have found that placing emphasis on either one of these attributes, i.e., MAF (Corbin et al., 2014; He et al., 2018) or LD (Badke et al., 2013; Ogawa et al., 2016), while maintaining an even distribution of SNP across the genome is more accurate than simply basing selection on an even distribution alone. He et al. (2018) achieved a 0.6 unit increase in the percentage correctly imputed SNP when a 6,000 SNP panel that maximized MAF was investigated for US Holstein cattle relative to selection solely to achieve and even distribution. In terms of LD as selection criteria, Ogawa et al. (2016) indicated a 0.02, 0.01, and 0.005 unit improvement in imputation accuracy (concordance rate) when their selection strategy included LD information, as opposed to even distribution alone, to construct 1,000, 4,000, and 10,000 SNP panels for Japanese Black cattle. The results presented in this study were therefore consistent with previous research on these selection strategies and suggests that both MAF and LD need to be considered in the design of a reduced panel for the SA Drakensberger and similar breeds.

The lowest SNP density that could be generated for the SA Drakensberger population using the LDCLUST method (23,469 SNP at the lowest LD threshold) surpassed all of the predefined SNP densities except the 50,000 SNP densities; this was a function of the mathematics underpinning the LDCLUST algorithm. Because the LDCLUST algorithm assigns

SNP to panels based on intra-autosomal LD, this was expected. Previous studies of SA Drakensberger have reported weak and variable autosome-wide LD (Lashmar et al., 2018), as well as LD of $r^2 \geq 0.2$ only persisting over short genomic distances between SNP (10 to 20kb; Makina et al., 2015). A larger number of clusters, and hence SNP density, that contain fewer strong-LD SNP was therefore expected for the admixed population in the present study in comparison to populations under intensive selection (e.g., layer chickens; Herry et al., 2018); no studies using such an approach in purebred cattle exists. While the LDCLUST method generated similar imputation accuracy to the best predefined selection method (MAFLD; Table 2), despite the fact that almost 20% more SNP were selected in the former, the LDCLUST method is expected to be of more value when higher density genotypes, or sequencing data, are available to better capture the true LD characteristics of an admixed breed such as the SA Drakensberger.

Chromosome-level variability

Several imputation studies in cattle have reported differences in imputation errors per chromosome (Berry and Kearney, 2011; Judge et al., 2016; Bernardes et al., 2018). Larger autosomes are expected to harbor more SNP, which facilitates the better capturing of stronger LD and subsequently enables more accurate inference of haplotypes (Sun et al., 2012). In agreement with previous research in cattle (Judge et al., 2016), imputation accuracy was best for the larger autosomes.

Differences in correlation-based imputation accuracy were observed for SNP located on the 2 autosomal extremes vs. in the middle of autosomes, and this was in agreement with studies such as Druet et al. (2010) and Badke et al. (2013). Badke et al. (2013), for example, demonstrated a 0.023 unit improvement in imputation accuracy for SNP in the middle vs. on the peripheries of chromosomes. Because the MAFLD selection methodology enriched the 2 extremes on each autosome with an extra SNP, higher mean imputation accuracy (+0.040 and +0.020 units higher than RAN for the first and last 1 Mb regions across chromosomes) was achieved for SNP located in these regions when this method was used. Furthermore, despite selecting additional SNP on the chromosomal extremities, Judge et al. (2016) still documented poorer imputation accuracy for shorter chromosomes due to the peripheral regions making up a larger proportion of the chromosome. Variation in SNP imputation accuracy within chromosomes was reduced in the present study as the SNP density of the panel increased suggesting that the effect of a chosen SNP location within chromosome is more detrimental to imputation accuracy when SNP density is low.

Minor allele frequency

The Sanga subspecies has been shown to suffer ascertainment bias when commercial genotyping panels have been used and this was evidenced by a higher proportion of low-MAF, less informative SNP, relative to taurine breeds (Qwabe et al., 2013; Zwane et al., 2016; Lashmar et al., 2018). This bias was introduced because these Sanga breeds were not included in the SNP discovery process involved during the original panel design. Although commercially available and custom SNP panels may have some utility for SA Sanga breeds, albeit sub-optimal, no Sanga-specific SNP panel exists at present (Zwane et al., 2019). The impact of low MAF on imputation accuracy has been extensively studied although the

documented impact varies depending on the parameter(s) used to define imputation accuracy (Mulder et al., 2012; Brøndum et al., 2014; Calus et al., 2014).

Concordance-based measures of accuracy (GCR_{SNP} and ACR_{SNP}) eroded with increasing MAF, whilst the correlation-based accuracy measure (COR_{SNP}) improved with increasing MAF. Correlation-based measures have been suggested to minimize the dependency of imputation accuracy on SNP allele frequency (Sargolzaei et al., 2014; Ventura et al., 2014). The difference in imputation accuracy between accuracy metrics was more pronounced for the RAN method than the MAFLD method as SNP were not specifically chosen to maximize MAF. The mean COR_{SNP} increased from 0.821 to 0.955 when averaged within low ($0.01 < MAF \leq 0.02$) vs. high ($0.4 < MAF \leq 0.5$) MAF classes. Similarly, Ma et al. (2013) documented an increase in the correlation coefficient from ~80% to ~85% when MAF increased from 0.1 to ~0.5 in European Red cattle.

The lower mean imputation accuracy for SNP with an MAF <10% suggests that imputation of rare variants was more challenging (Calus et al., 2014), which is concerning considering that these variants may be associated with unique or complex traits such as those pertaining to adaptability (Zwane et al., 2019).

Conclusion

Genotyping at higher densities is costly and currently financially not feasible within beef industries of the developing world. Genotyping selection candidates on lower densities and imputing to higher density should reduce the costs involved in routinely applying GS for locally important breeds with unique adaptation abilities. Results from this study indicated that a custom genotyping panel consisting of 10,000 SNP would suffice in achieving <3% imputation errors for the SA Drakensberger. The reduced-density SNP should be selected based on criteria such as MAF and LD that are specific to the breed to maximize achievable imputation accuracy. Whole-genome sequencing data would, however, serve as a better resource of candidate SNP to select from in terms of abundance and ensuring breed specificity of SNP. The inferences made in this study for the SA Drakensberger may be transferable to other Sanga breeds and may serve as guidelines in future imputation-driven genomic endeavors for these breeds and other breeds that have admixed genomes.

Acknowledgments

This project received funding from both the Beef Genomics Project (BGP; grant number: BGP15003) and Red Meat Research and Development South Africa (RMRDSA), and the authors would like to thank these funding bodies for their financial support.

Funding

This article received funding from Red Meat Research and Development South Africa (RMRDSA) and the Beef Genomics Program.

Conflict of Interest Statement

The authors declare no real or perceived conflicts of interest.

Literature Cited

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. doi:10.3168/jds.2018-14621

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, J. Fix, C. P. Van Tassell, and J. P. Steibel. 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. *BMC Genet*. 14:8. doi:10.1186/1471-2156-14-8

Bernardes, P. A., H. A. Al-Mamun, M. Suarez, D. Lim, B. Park, and C. Gondro. 2018. Imputation accuracy of whole-genome sequence data in Hanwoo cattle. *Proc. World. Cong. Genet. Appl. Livest. Prod.* 11.

Berry, D. P., and J. F. Kearney. 2011. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. *Animal* 5:1162–1169. doi:10.1017/S1751731111000309

Brøndum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC Genomics* 15:728. doi:10.1186/1471-2164-15-728

Calus, M. P., 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 applications. *Animal* 8:1743–1753.doi:10.1017/S1751731114001803

Carvalheiro, 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 Nellore cattle. *Genet. Sel. Evol.* 46:69. doi:10.1186/s12711-014-0069-1

Corbin, L. J., A. Kranis, S. C. Blott, J. E. Swinburne, M. Vaudin, S. C. Bishop, and J. A. Woolliams. 2014. The utility of low-density genotyping for imputation in the Thoroughbred horse. *Genet. Sel. Evol.* 46:9. doi:10.1186/1297-9686-46-9

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

Ferraz, J. B. S., X. Wu, H. Li, J. Xu, R. Ferretti, B. Simpson, J. Walker, L. R. Silva, J. F. Garcia, R. G. Tait Jr., and S. Bauck. 2018. Design of a low-density SNP chip for Bos indicus: GGP indicus technical characterization and imputation accuracy to higher density SNP genotypes. Proceedings of the World Congress on Genetics Applied to Livestock Production. Auckland, New Zealand; p. 6–11.

García-Ruiz, A., F. J. Ruiz-Lopez, G. R. Wiggans, C. P. Van Tassell, and H. H. Montaldo. 2015. Effect of reference population size and available ancestor genotypes on imputation of Mexican Holstein genotypes*. J. Dairy Sci.* 98:3478–3484. doi:10.3168/jds.2014-9132

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. van der Werf. 2012. Accuracy of genotype imputation in sheep breeds. *Anim. Genet*. 43:72–80. doi:10.1111/j.1365-2052.2011.02208.x

He, J., J. Xu, X. L. Wu, S. Bauck, J. Lee, G. Morota, S. D. Kachman, and M. L. Spangler. 2018. Comparing strategies for selection of low-density SNPs for imputation-mediated genomic prediction in U. S. Holsteins. *Genetica* 146:137–149. doi:10.1007/s10709-017-0004-9

Hérault, F., J. Yon, F. Herry, S. Allais, and P. Le Roy. 2016. SS4I: select SNP subset for imputation. 2016 (in French). Available from https://prodinra.inra.fr/record/375448.

Herry, F., F. Hérault, D. Picard Druet, A. Varenne, T. Burlot, P. Le Roy, and S. Allais. 2018. Design of low density SNP chips for genotype imputation in layer chicken. *BMC Genet*. 19:108. doi:10.1186/s12863-018-0695-7

Judge, M. M., J. F. Kearney, M. C. McClure, R. D. Sleator, and D. P. Berry. 2016. Evaluation of developed low-density genotype panels for imputation to higher density in independent dairy and beef cattle populations1. *J. Anim. Sci.* 94:949–962. doi: 10.2527/jas.2015-0044

Júnior, G. A. O., 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, et al. 2017. Genotype imputation in a tropical crossbred dairy cattle population. *J. Dairy Sci.* 100:9623–9634. doi:10.3168/jds.2017-12732

Kaufman, L., and P. Rousseeuw. 2009. *Finding groups in data*. Hoboken: John Wiley and Sons.

Lashmar, S. F., C. Visser, E. van Marle-Köster, and F. C. Muchadeyi. 2018. Genomic diversity and autozygosity within the SA Drakensberger beef cattle breed. *Livest. Sci*. 212:111–119. doi: 10.1016/j.livsci.2018.04.006

Ma, P., R. F. Brøndum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *J. Dairy Sci.* 96:4666–4677. doi:10.3168/jds.2012-6316

Makina, S. O., J. F. Taylor, E. van Marle-Köster, F. C. Muchadeyi, M. L. Makgahlela, M. D. MacNeil, and A. Maiwashe. 2015. Extent of linkage disequilibrium and effective population size in four South African Sanga cattle breeds. *Front. Genet*. 6:337. doi:10.3389/fgene.2015.00337

Makina, S. O., L. K. Whitacre, J. E. Decker, J. F. Taylor, M. D. MacNeil, M. M. Scholtz, E. van Marle-Köster, F. C. Muchadeyi, M. L. Makgahlela, and A. Maiwashe. 2016. Insight into the genetic composition of South African Sanga cattle using SNP data from cattle breeds worldwide. *Genet. Sel. Evol*. 48:88. doi:10.1186/s12711-016-0266-1

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. Smith, T. S. Sonstegard, et al. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* 4:e5350. doi:10.1371/journal.pone.0005350

Mulder, H. A., M. P. 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

Nicolazzi, E. L., S. Biffani, F. Biscarini, P. Orozco Ter Wengel, A. Caprera, N. Nazzicari, and A. Stella. 2015. Software solutions for the livestock genomics SNP array revolution. *Anim. Genet*. 46:343–353. doi:10.1111/age.12295

Nicolazzi, E. L., G. Marras, and A. Stella. 2016. SNPConvert: SNP array standardization and integration in livestock species. *Microarray*s 5:17. doi: 10.3390/microarrays5020017

O'Brien, A. C., M. M. Judge, S. Fair, and D. P. Berry. 2019. High imputation accuracy from informative low-to-medium density single nucleotide polymorphism genotypes is achievable in sheep1. J. *Anim. Sci*. 97:1550–1567. doi:10.1093/jas/skz043

Ogawa, S., H. Matsuda, Y. Taniguchi, T. Watanabe, A. Takasuga, Y. Sugimoto, and H. Iwaisaki. 2016. Accuracy of imputation of single nucleotide polymorphism marker genotypes from low-density panels in Japanese Black cattle. *Anim. Sci.* J. 87:3–12. doi:10.1111/asj.12393

Qwabe, S. O., E. VanMarle-Köster, A. Maiwashe, and F. C. Muchadeyi. 2013. Short communication: evaluation of the BovineSNP50 genotyping array in four South African cattle populations. *S. Afr. J.Anim. Sci.* 43: 64–67. doi:10.4314/sajas.v43i1.7

SA Drakensberger Breeders' Society. 2011. *Drakensberger Handbook* (1st Ed.). Available from http://www.drakensbergers.co.za/pdf/manual.pdf.

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

Shifman, S., J. Kuypers, M. Kokoris, B. Yakir, and A. Darvasi. 2003. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet*. 12:771–776. doi:10.1093/hmg/ddg088

Sun, C., X. L. Wu, K. A. Weigel, G. J. 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. doi:10.1017/S001667231200033X

Toosi, A., R. L. Fernando, and J. C. Dekkers. 2010. Genomic selection in admixed and crossbred populations. *J. Anim. Sci*. 88:32–46. doi:10.2527/jas.2009-1975

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 (Bethesda*). 7:1377–1383. doi:10.1534/g3.117.040717

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 cattle1. *J. Anim. Sci*. 92:1433–1444. doi: 10.2527/jas.2013-6638

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

Zwane, A. A., A. Maiwashe, M. L. Makgahlela, A. Choudhury, J. F. Taylor, and E. Van Marle-Köster. 2016. Genome-wide identification of breed-informative single-nucleotide polymorphisms in three South African indigenous cattle breeds. *S. Afr. J. Anim. Sci*. 46:302–312. doi: 10.4314/sajas.v46i3.10

Zwane, A. A., R. D. Schnabel, J. Hoff, A. Choudhury, M. L. Makgahlela, A. Maiwashe, E. Van Marle-Koster, and J. F. Taylor. 2019. Genome-wide SNP discovery in indigenous cattle breeds of South Africa. *Front. Genet*. 10:273. doi:10.3389/fgene.2019.00273