J. Dairy Sci. 98:4969–4989
http://dx.doi.org/10.3168/jds.2014-9213

Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (*Bos indicus*) dairy cattle: Comparison of commercially available SNP chips

S. A. Boison,*[1] D. J. A. Santos,† A. H. T. Utsunomiya,† R. Carvalheiro,† H. H. R. Neves,† A. M. Perez O'Brien,* J. F. Garcia,‡ J. Sölkner,* and M. V. G. B. da Silva§
*University of Natural Resources and Life Sciences, Department of Sustainable Agricultural Systems, Gregor-Mendel 33, A-1180, Vienna, Austria
†Faculdade de Ciências Agrária e Veterinárias, Universidade Estadual Paulista (UNESP), SP, 148841900, Brazil
‡Faculdade de Medicina Veterinária de Araçatuba, Universidade Estadual Paulista (UNESP), Araçatuba, SP, 16015-050, Brazil
§Empresa Brasileira de Pesquisa Agropecuária, Embrapa Gado de Leite, Juiz de Fora, MG, 36038-330, Brazil

ABSTRACT

Genotype imputation is widely used as a cost-effective strategy in genomic evaluation of cattle. Key determinants of imputation accuracies, such as linkage disequilibrium patterns, marker densities, and ascertainment bias, differ between *Bos indicus* and *Bos taurus* breeds. Consequently, there is a need to investigate effectiveness of genotype imputation in indicine breeds. Thus, the objective of the study was to investigate strategies and factors affecting the accuracy of genotype imputation in Gyr (*Bos indicus*) dairy cattle. Four imputation scenarios were studied using 471 sires and 1,644 dams genotyped on Illumina BovineHD (HD-777K; San Diego, CA) and BovineSNP50 (50K) chips, respectively. Scenarios were based on which reference high-density single nucleotide polymorphism (SNP) panel (HDP) should be adopted [HD-777K, 50K, and GeneSeek GGP-75Ki (Lincoln, NE)]. Depending on the scenario, validation animals had their genotypes masked for one of the lower-density panels: Illumina (3K, 7K, and 50K) and GeneSeek (SGGP-20Ki and GGP-75Ki). We randomly selected 171 sires as reference and 300 as validation for all the scenarios. Additionally, all sires were used as reference and the 1,644 dams were imputed for validation. Genotypes of 98 individuals with 4 and more offspring were completely masked and imputed. Imputation algorithms FImpute and Beagle v3.3 and v4 were used. Imputation accuracies were measured using the correlation and allelic correct rate. FImpute resulted in highest accuracies, whereas Beagle 3.3 gave the least-accurate imputations. Accuracies evaluated as correlation (allelic correct rate) ranged from 0.910 (0.942) to 0.961 (0.974) using 50K as HDP and with 3K (7K) as low-density panels. With GGP-75Ki as HDP, accuracies were moderate for 3K, 7K, and 50K, but high for SGGP-20Ki. The use of HD-777K as HDP resulted in accuracies of 0.888 (3K), 0.941 (7K), 0.980 (SGGP-20Ki), 0.982 (50K), and 0.993 (GGP-75Ki). Ungenotyped individuals were imputed with an average accuracy of 0.970. The average top 5 kinship coefficients between reference and imputed individuals was a strong predictor of imputation accuracy. FImpute was faster and used less memory than Beagle v4. Beagle v4 outperformed Beagle v3.3 in accuracy and speed of computation. A genotyping strategy that uses the HD-777K SNP chip as a reference panel and SGGP-20Ki as the lower-density SNP panel should be adopted as accuracy was high and similar to that of the 50K. However, the effect of using imputed HD-777K genotypes from the SGGP-20Ki on genomic evaluation is yet to be studied.

Key words: imputation, Gyr, Beagle, FImpute

INTRODUCTION

Genomic selection (**GS**) is now implemented by many livestock- and plant-breeding organizations. In general, genotyping chips with approximately 50 to 60 thousand SNP have been extensively used in livestock populations for genomic evaluation (sheep: Daetwyler et al., 2012a,b; pigs: Cleveland and Hickey, 2013; Badke et al., 2014; cattle: Harris and Johnson, 2010; Brøndum et al., 2011). Studies on increasing the accuracy of genomic EBV using approximately 800,000 SNP markers have also been reported (e.g., Erbe et al., 2012). Additionally, the use of lower-density SNP panels (few hundreds to about 7,000) has been studied (Weigel et al., 2010a; Vazquez et al., 2010; Dassonneville et al., 2012). Compared with traditional breeding methods,

Received December 9, 2014.
Accepted March 22, 2015.
[1]Corresponding author: solomon.boison@students.boku.ac.at

higher response to selection has been reported for genomic evaluations, especially in dairy cattle (Hutchison et al., 2014).

An important issue with the implementation of GS is the cost of genotyping individuals (i.e., reference and selection candidates). Although the last decade has witnessed a drastic reduction in genotyping cost, the current price is still relatively high. To reduce the cost of implementing GS, imputation strategies have been suggested and explored (Huang et al., 2012; Khatkar et al., 2012). One key strategy is to genotype reference individuals with a high-density panel (**HDP**) and selection candidates with a lower-density marker panel (**LDP**). Untyped markers of the HDP are then inferred for the individuals genotyped with the LDP. Imputation is also used to combine marker information from different genotyping platforms (Howie et al., 2009).

Broadly, imputation methods can be classified into 2 groups: (a) those based on population linkage disequilibrium (**LD**) information [e.g., Impute2 (Howie et al., 2009), Beagle (Browning and Browning, 2007, 2013), MaCH (Li et al., 2010)] and (b) those based on family, segregation, or a combination of family, segregation, and population LD [e.g., AlphaImpute (Hickey et al., 2012b), Findhap (VanRaden et al., 2011), DAGPHASE (Druet and Georges, 2010), FImpute (Sargolzaei et al., 2008, 2014), PedImpute (Nicolazzi et al., 2013)]. For population-based methods, Impute2 has been shown to be highly accurate in cattle populations (Ma et al., 2013; Brøndum et al., 2014; Sargolzaei et al., 2014) and Beagle has been extensively used. Family or segregation rule-based methods are known to be fast and can be as accurate as population-based algorithms, especially when individuals genotyped with LDP have close relatives in the reference population (Hickey et al., 2012b; Sargolzaei et al., 2014).

Accuracy of imputation is affected by several factors, for example number and composition of reference individuals used to build the haplotype library, effective population size, marker allele frequencies, and difference between marker densities of the reference and imputed sets (Badke et al., 2013; Ma et al., 2013; Brøndum et al., 2014; Sargolzaei et al., 2014). The effect of these factors on imputation accuracy have been widely explored in *Bos taurus* cattle breeds (Berry and Kearney, 2011; Khatkar et al., 2012; Ma et al., 2013; Pausch et al., 2013; Larmer et al., 2014).

With LD and minor allele frequency patterns differing between *Bos taurus* and *Bos indicus* breeds (Pérez O'Brien et al., 2014), studies should be conducted to ascertain imputation effectiveness in *Bos indicus* breeds. Furthermore, the performance of different marker density panels (either as LDP or HDP) and of different genotyping platforms need to be investigated. For example, the Illumina BovineSNP50 (Illumina Inc., San Diego, CA) SNP chip is known to have high ascertainment bias toward taurine breeds (Utsunomiya et al., 2014). GeneSeek (GeneSeek Inc., Lincoln, NE) has introduced SNP panels (SGGP-20Ki and GGP-75Ki) claiming to be specific for *Bos indicus* breeds and to address the issue of high ascertainment bias of the Illumina BovineSNP50. A recent study by Carvalheiro et al. (2014) on Nelore (*Bos indicus*) beef cattle suggests that imputation accuracies could be similar to those reported for *Bos taurus* breeds.

The objective of the current study was to investigate strategies and factors affecting the accuracy of genotype imputation in Gyr dairy cattle, comparing imputation accuracies obtained from different combinations of HDP (Illumina HD-777K, GeneSeek GGP-75Ki, and Illumina 50K) and LDP (Illumina 3K, Illumina 7K, GeneSeek SGGP-20Ki, Illumina 50K, GeneSeek GGP-75Ki). Imputation accuracy of ungenotyped individuals was also studied. Population-based imputation (Beagle version 3.3 and newly released version 4; Browning and Browning, 2007, 2013) and population plus linkage-based methods (FImpute; Sargolzaei et al., 2014) were tested.

MATERIALS AND METHODS

Study Samples

The samples used in this study are from Gyr (*Bos* indicus) dairy cattle. All genotyped animals are from a breeding program in Brazil, where Embrapa Dairy Cattle (Jiuz de Fora, Brazil) is responsible for genetic evaluations. The Gyr breed has about 4,000 sires registered in the herd book of the breed association; however, DNA samples are not available for most of them. The samples available for the present study consisted of 474 bulls, genotyped with the Illumina BovineHD (HD-777K), and 1,688 cows genotyped with the Illumina BovineSNP50 v2 (50K).

Pedigree information of the genotyped individuals after following the genotype quality check is described as follows. There were 5,153 animals (out of which 978 were founders) in the pedigree. The pedigree also consisted of 871 sires and 2,736 dams. Individuals with only 1 known sire or dam and with both parents known were 442 and 3,734, respectively.

Genotype Quality Checks

Genotype quality checks were done on the samples genotyped with the HD-777K using PLINK v1.07 (Purcell et al., 2007). The SNP markers with call rate <90%, minor allele frequency (**MAF**) <0.02, and *P*-

value for the Hardy-Weinberg equilibrium test $<10^{-6}$ were discarded. The SNP with unknown and same physical position were also removed. Only autosomal SNP were used. Samples with call rate <90% were removed. Parent-offspring pedigree relationships were set to missing if more than 180 SNP (value determined based on empirical distribution of Mendelian errors) were detected inconsistent based on Mendelian segregation rules (opposing homozygotes between parent and offspring). After setting parent-offspring pedigree relationships to unknown, locus-specific inconsistent genotypes of the remaining parent-offspring pairs were set to missing. The other SNP panels used in our study were subsets of the HD-777K panel. Quality checks of SNP data for the dams was done to discard SNP and samples with poor call rate (samples and SNP with call rate <90% were all discarded). Details about the number of markers and samples after quality check are presented in Table 1. See Supplementary Table S1 (http://dx.doi.org/10.3168/jds.2014-9213) for details about marker data before quality check.

Imputation Scenarios

Three main scenarios were studied, considering different combinations of HDP (HD-777K, GGP-75Ki and 50K) and LDP (3K, 7K, 50K, SGGP-20Ki and GGP-75Ki) for the reference and imputed sets, respectively. An additional scenario on imputing ungenotyped individuals was also undertaken.

Scenario 1 was based on using the 50K as the HDP for the reference set, and the 3K and 7K as the LDP for the imputed set. In scenario 2, the HD-777K was used as the HDP, and the 3K, 7K, 50K, SGGP-20Ki, and GGP-75Ki as the LDP. Scenario 3 used GGP-75Ki as the HDP and 3K, 7K, SGGP-20Ki, and 50K as the LDP. For scenario 4, we masked the genotypes (setting them to ungenotyped state) of 98 sires and dams with 4 and more genotyped progeny and imputed them. More information about the scenarios is presented in Table 1.

Different reference and imputed sets were tested. Initially, the sires were randomly split in 2 groups, the reference comprising 171 sires and the imputed set with 300 sires, (a ratio of about 1:2). The true genotypes of the individuals in the imputed set were masked, except for the SNP present in the LDP, and then imputed. An alternate analysis was done with a reference-to-imputed set ratio of ∼2:1 to assess the gain in imputation accuracy when using a larger reference population. As only the sires were genotyped with the HD-777K, they were used in all the scenarios described above. The dams were genotyped with the 50K and were only used in scenario 1. The dams had their genotypes masked, except for the SNP present in the LDP (3K and 7K), and then imputed using all the genotyped sires as the reference set.

Genotype Imputation

Genotype imputation of untyped marker loci for all scenarios was undertaken using Beagle v4 release 1182 (Browning and Browning, 2013), Beagle v3.3 (Browning and Browning, 2007), and FImpute v2.2 (Sargolzaei et al., 2014). Beagle v4 was used with default parameters

Table 1. Number of samples and SNP after quality check for all the tested scenarios

Scenario	Chip[1]	No. of reference/imputed[2]	No. of SNP in the low-density panel	No. of SNP in the high-density panel	SNP to be imputed (%)
1	LD3K_50K	471/1,644	1,874	22,152	91.5
	LD7K_50K	471/1,644	4,239	22,152	80.9
	LD3K _50K	171/300	1,874	22,152	91.5
	LD7K _50K	171/300	4,239	22,152	80.9
	LD3K _HD	171/300	2,138	496,606	99.6
	LD7K_HD	171/300	4,727	496,606	99.0
2	LD20Ki_HD	171/300	15,727	496,606	96.8
	LD50Ki_HD	171/300	22,152	496,606	95.0
	LD75Ki_HD	171/300	65,012	496,606	86.9
	LD3K_75Ki	171/300	1,869	65,018	97.1
3	LD7K_75Ki	171/300	4,730	65,018	92.7
	LD20Ki_75Ki	171/300	13,854	65,018	78.7
	LD50K_75Ki	171/300	7,239	65,018	88.9
4	LD0K_50K	1,970/98	0	22,152	100.0

[1]SNP chips for low- and high-density panels; HD = Illumina BovineHD (Illumina, San Diego, CA); 50K = Illumina BovineSNP50; 3K = Illumina Bovine3K; 7K = Illumina BovineLD; 20Ki = GeneSeek SGGP Indicus LD (Geneseek, Lincoln, NE); 75Ki = GeneSeek GGP Indicus HD.

[2]Reference animals were all sires (n = 171 or n = 471), and imputed animals were either sires (n = 300) or dams (1,644) for scenarios 1, 2, and 3. Scenario 4 consisted of both sires and dams as reference (n = 1,970) and imputed (n = 98) animals.

(increasing phasing run and number of iteration did not increase accuracy in this study), and the older version Beagle v3.3 was also tested to study differences in imputation accuracy as well as in terms of computation time. For both Beagle v3.3 and Beagle v4, we performed imputation by first phasing HDP individuals separately (prephasing with 10 iterations). Subsequently, phased HDP individuals were used to impute LDP individuals; LDP individuals were not prephased. According to Browning and Browning (2013), Beagle v4 should be better in phasing and inferring untyped markers than Beagle v3.3.

FImpute has been implemented to use pedigree information to unambiguously phase and infer untyped markers for close relatives using an iterative approach. When pedigree information is not available, long and short sliding windows are built from the population data to help identify shared haplotypes for accurate phasing and imputation. Imputation with FImpute was undertaken with and without pedigree information, to enable comparison with Beagle.

Imputation Accuracy: Sample- and SNP-Specific Accuracy

Imputation accuracies were based on genotypes (FImpute and Beagle) and allele dosages (Beagle). Accuracies were computed using custom scripts in the statistical software package R (R Development Core Team, 2011).

Sample-specific imputation accuracies were calculated using the 2 measures described below. Each measure has its distinct feature, as some are able to reduce the effect of allele frequency on accuracy of imputation.

1) One minus the mean of absolute difference between observed and imputed genotype ($\mathbf{ACR_{anim}}$). It is a more relaxed measure of imputation accuracy than percent of correct genotypes, as imputation of homozygote to heterozygote (or vice versa) would only count as half error. Additionally, it does not penalize for easy imputation of markers with low MAF. $\mathrm{ACR_{anim}}$ was computed as

$$\mathrm{ACR_{anim}} = 1 - \frac{\Sigma_{j=1}^{L_k} \mid g_{jk} - \hat{g}_{jk} \mid}{2 \times L},$$

where L represents the total number of markers, g_{jk} is the observed (true) genotype (0, 1, 2) for SNP j of individual k, $\hat{g}_{jk}$ is the imputed genotype (0, 1, 2) for SNP j of individual k.

2) Pearson correlation between observed and imputed genotype ($\mathbf{corr_{anim}}$). Its main advantage is the robustness to the effect of MAF on the measure of imputation accuracy:

$$\mathrm{corr_{anim}} = \frac{\Sigma_{j=1}^{L_k} \left(g_{jk} - \overline{g}\right)\left(\hat{g}_{jk} - \overline{\hat{g}}\right)}{\sqrt{\Sigma_{j=1}^{L_k} \left(g_{jk} - \overline{g}\right)^2 \Sigma_{j=1}^{L_k} \left(\hat{g}_{jk} - \overline{\hat{g}}\right)^2}},$$

where L, k, j, g_{jk}, and $\hat{g}_{jk}$ are as defined under $\mathrm{ACR_{anim}}$; $\overline{g}$ and $\overline{\hat{g}}$ denote the average value of observed and imputed genotypes, respectively.

The SNP-specific imputation accuracies were computed for each SNP as $\mathbf{ACR_{snp}}$ and $\mathbf{corr_{snp}}$. The formulas to compute SNP-specific accuracies is similar to that of the sample-specific accuracies except that for $\mathrm{ACR_{snp}}$ and $\mathrm{corr_{snp}}$, L_k and $j = 1$ in $\mathrm{ACR_{anim}}$ and $\mathrm{corr_{anim}}$ are replaced with N_j and $k = 1$, respectively; N_j denotes the total number of animals for SNP j.

After computing the SNP-specific accuracies, the effect of MAF on imputation accuracy was also assessed (Badke et al., 2013; Hickey et al., 2012a). Minor allele frequencies were binned at 0.01 intervals for graphical presentation. In addition, SNP-specific accuracies were computed after deleting markers that were predicted by Beagle to potentially have low imputation accuracy. Beagle internally predicts imputation accuracy of an SNP (allelic R^2) as the squared correlation between the most likely allele dosage and the true allele dosage. True allele dosages were estimated using posterior genotype probabilities. The possibility of increasing the overall imputation accuracy, at the expense of losing some markers, was also investigated. This was done by discarding SNP with allelic R^2 lower than a specified threshold (0.7, 0.8, 0.85, and 0.90) and imputation accuracy was calculated for the SNP left.

Intermarker Distances, LD, and Marker Density at Different MAF Intervals

The average intermarker distances, LD, and allele frequency distribution were computed to provide information about each SNP chip used for this study. Linkage disequilibrium was calculated as the squared correlation (r^2) between all marker pairs with PLINK v1.07 (Purcell et al., 2007) software; LD was binned at 10-kb intervals and graphical presentations are made to illustrate the trends observed for each SNP chip.

Relatedness Between Reference and Imputed Sets

The effect of having relatives in the reference population for imputation was studied. Using the 50K geno-

type data which was available for all animals and the approach of VanRaden (2008), genomic relationships (G) were estimated:

$$G = \frac{\mathbf{ZZ}'}{2\Sigma_{i=1}^{Nsnps} p_i\left(1-p_i\right)},$$

where $\mathbf{Z} = \mathbf{M} - \mathbf{P}$; $\mathbf{M}$ is a matrix of dimension number of animals × number of SNP markers (*Nsnps*) coded as (−1, 0, 1); $\mathbf{P}$ is a matrix containing the allele frequency expressed as a difference from 0.5 and multiplied by 2: $\mathbf{P} = 2p_i - 1 = 2(p_i - 0.5)$; p_i is the allele frequency at a locus estimated from the observed genotype data.

Four statistics were computed based on the genomic relationships between each imputed animal and the animals in the reference set: maximum relationship (**relmax**), the average of top 5 relationships (**rel5**), the average of top 10 relationships (**rel10**), and the average of all relationships > 0 (**overall+**) relationship to the reference set. In addition, results of eigen decomposition of the kinship matrix were used to create plots representing the structure of our data set. The relationship between relatedness and imputation accuracy was assessed by regressing imputation accuracy on relatedness applying a second-order polynomial model.

Computing Time

Computation efficiency based on computing time and memory usage was computed for each scenario on using FImpute, Beagle v4, and Beagle v3.3. Due to the possibility of parallel computing all chromosomes at the same time, results are shown for the largest chromosome (BTA 1). All computations were undertaken on an Intel Xeon CPU E5540 at 2.53GHz with 24.7 GB of RAM (Intel, Santa Clara, CA).

RESULTS

Summary of Intermarker Distances and Marker Density at Different MAF Intervals

Figure 1 shows the distribution of MAF for each marker panel. The Illumina SNP chips (except 3K) had the highest proportion of markers with low MAF. The opposite was observed for the GeneSeek SNP chips, as most markers were highly polymorphic (Figure 1). It is important to note that these proportions were computed after deleting markers with MAF <0.02. About 13.3 to 17.4% of the total number of markers had MAF <0.05 for HD-777K, 50K, and 7K. The value increased to about 34.6% for MAF <0.10. Alternatively, 4.6 to 6.0% of the total number of markers on the SGGP-20Ki and GGP-75Ki had MAF <0.05. The value increased to 11.8 to 13.5% for MAF <0.10. However, the cumulative number of markers between MAF 0.10 and 0.40 tended to be similar for all the SNP chips (Figure 1). The average (median) allele frequencies computed with the genotyped sires for 3K, 7K, SGGP-20Ki, 50K, GGP-75Ki, and HD-777K were 0.255 (0.253), 0.214 (0.190), 0.306 (0.340), 0.199 (0.168), 0.299 (0.320), and 0.221 (0.200), respectively.

LD Patterns

The trends of LD decay for HD-777K, 50K, GGP-75Ki, and SGGP-20Ki were similar (Figure 2). Gradual decrease in LD over genomic distance was observed. Figure 2 also shows r^2 at the average intermarker distances. The average r^2 for the HD-777K at an average intermarker distance of ∼5kb was 0.378. The r^2 of 0.210, 0.138, 0.105, 0.056, and 0.047 at average intermarker distance of ∼39, ∼100, ∼160, ∼528, and ∼1,153 kb were observed for GGP-75Ki, 540K, SGGP-20Ki, 7K, and 3K respectively. In addition, the average (median) r^2 of adjacent (syntenic) markers for the 6 marker panels were 0.052 (0.020) for 3K, 0.085 (0.025) for 7K, 0.131 (0.052) for SGGP-20Ki, 0.166 (0.048) for 50K, 0.237 (0.125) for GGP-75Ki, and 0.426 (0.282) for HD-777K.

Relatedness Between Reference and Imputed Sets

Table 2 shows the average relatedness between reference individuals as well as between reference and imputed individuals. The maximum estimated genomic relationship between the reference and imputed set suggests that most animals had their sire or full-sib in the reference population. The degree of relatedness was higher between sires (reference) and dams (imputed) than within the genotyped sire population. The increase of reference individuals to 300 for the sire to sire scenario resulted in an increase in average relatedness. Relatedness estimates of about 0.24 to 0.28 and 0.19 to 0.23 between reference and imputed set for rel5 and rel10, respectively, imply that most individuals had at least a genotyped half-sib in the reference population (Table 2). Using LDP (3K, 7K, SGGP-20Ki) to estimate genomic relatedness was similar to that of using the 50K (results not shown). For the sire (n = 171)-to-sire (n = 300) scenario of ∼1:2, 207 out of the 300 imputed individuals had a genotyped sire in the reference population. Additionally for the sire (n = 471)-to-dam (n = 1644) scenario, 1,203 dams had a genotyped sire in the reference population. The direct visualization of the selected reference and imputed sets can be seen in

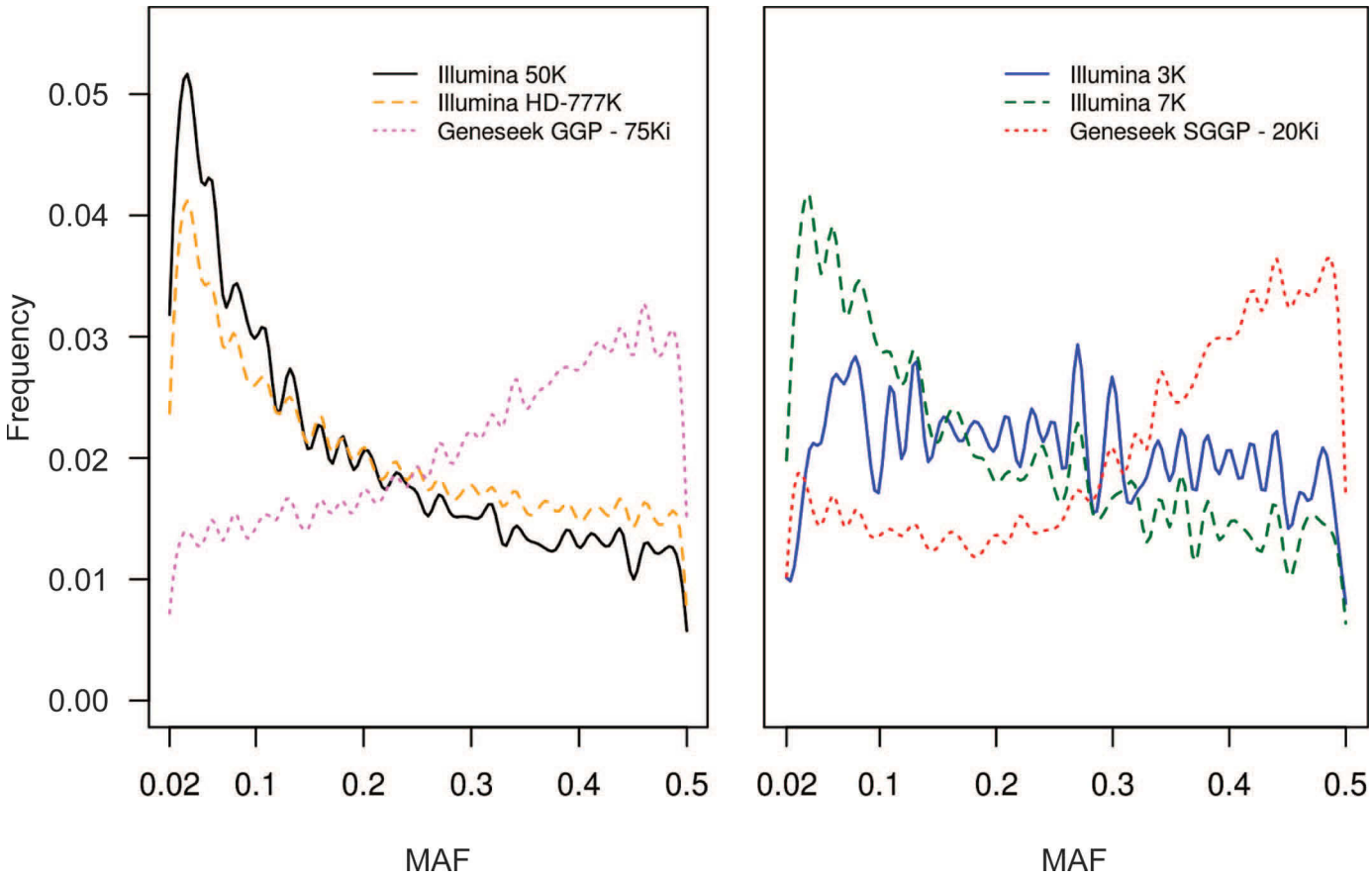

Figure 1. Distribution of minor allele frequency (MAF) for different SNP chips; MAF was binned at 0.01 intervals (Illumina, San Diego, CA; Geneseek, Lincoln, NE). Color version available online.

Figure 3 and the distribution of each of the relatedness measure (relmax, rel5, and rel10) can be seen in Supplementary Figure S1 (http://dx.doi.org/10.3168/jds.2014-9213). The graph shows the first 2 principal components estimated based on genomic kinship. The 2 principal components explained about 6.23% of the total variation and the graph corroborate the results of Table 2, indicating that reference animals are well nested within the imputed set.

Imputation Accuracy

Generally, sample- and SNP-specific accuracies were highest for FImpute followed by Beagle v4 and v3.3 in all the scenarios studied (Tables 3 and 4 and Figure 4). Accuracy increased with increasing marker density and with higher reference population size. The rank of individuals as well as the rank of the software packages did not change regardless of the measure of imputation accuracy used.

Sample-Specific Accuracies

Imputation of LDP (3K and 7K) to HDP (50K). Average imputation accuracies for scenario 1 (imputation of LDP 3K and 7K to 50K) are presented in Table 3. Results are presented for Beagle v3.3 and v4, as well as for FImpute with and without pedigree information. In all the scenarios, FImpute without pedigree information was at least as accurate as with pedigree information and was also the most accurate in some cases. Accuracies were higher for sire (reference; n = 471)-to-dam (imputed; n = 1,644) scenario than from the sire (reference; n = 171)-to-sire (imputed, n = 300) scenario. The difference in accuracy between the sire-to-sire and sire-to-dam scenarios can be explained by the highest relatedness between the dams and sires (Table 2) and the larger reference size (3 times) of the sire-to-dam scenario. Accuracies were also higher using LDP 7K than using 3K (Table 3).

Increasing the reference population increased sample-specific accuracy. Comparing the same individuals (n

Table 2. Average (SD) genomic kinship/relationships for animals within the reference data set and between animals in the imputed and reference data set

		Genomic kinship[2]			
Reference/imputed[1]	Information	relmax	rel5	rel10	Overall$^+$
171/300	Within reference	0.351 (0.127)	0.231 (0.072)	0.183 (0.062)	0.060 (0.022)
	Between imputed and reference	0.350 (0.122)	0.238 (0.071)	0.192 (0.062)	0.063 (0.025)
300/171	Within reference	0.433 (0.111)	0.302 (0.085)	0.249 (0.078)	0.071 (0.030)
	Between imputed and reference	0.395 (0.124)	0.258 (0.080)	0.205 (0.068)	0.058 (0.024)
471/1644	Within reference	0.451 (0.099)	0.318 (0.076)	0.263 (0.069)	0.065 (0.027)
	Between imputed and reference	0.433 (0.113)	0.280 (0.070)	0.228 (0.062)	0.057 (0.024)

[1]Reference animals were all sires (n = 171, n = 300, n = 471), and imputed animals were either sires (n = 300, n = 171) or dams (1,644).
[2]Maximum (relmax), average of top 5 (rel5) and top 10 (rel10) relationships, and the average overall relationship ≥0 (Overall$^+$) of imputed or reference individual to the reference set.

= 171) imputed for the sire to sire scenario with 171 reference individuals to a scenario with 300 reference individuals, $corr_{anim}$ increased by 0.016 and 0.010 correlation points for 3K and 7K with FImpute. Higher increases were also observed with Beagle v4 and v3.3 (results not shown).

Imputation of LDP (3K, 7K, SGGP-20Ki, 50K, and GGP-75Ki) to HDP (HD-777K). As was observed for scenario 1 (imputation of LDP 3K and 7K to HDP 50K), accuracy increased with increasing marker density; FImpute was still the most accurate software (Figure 4) and increasing reference population

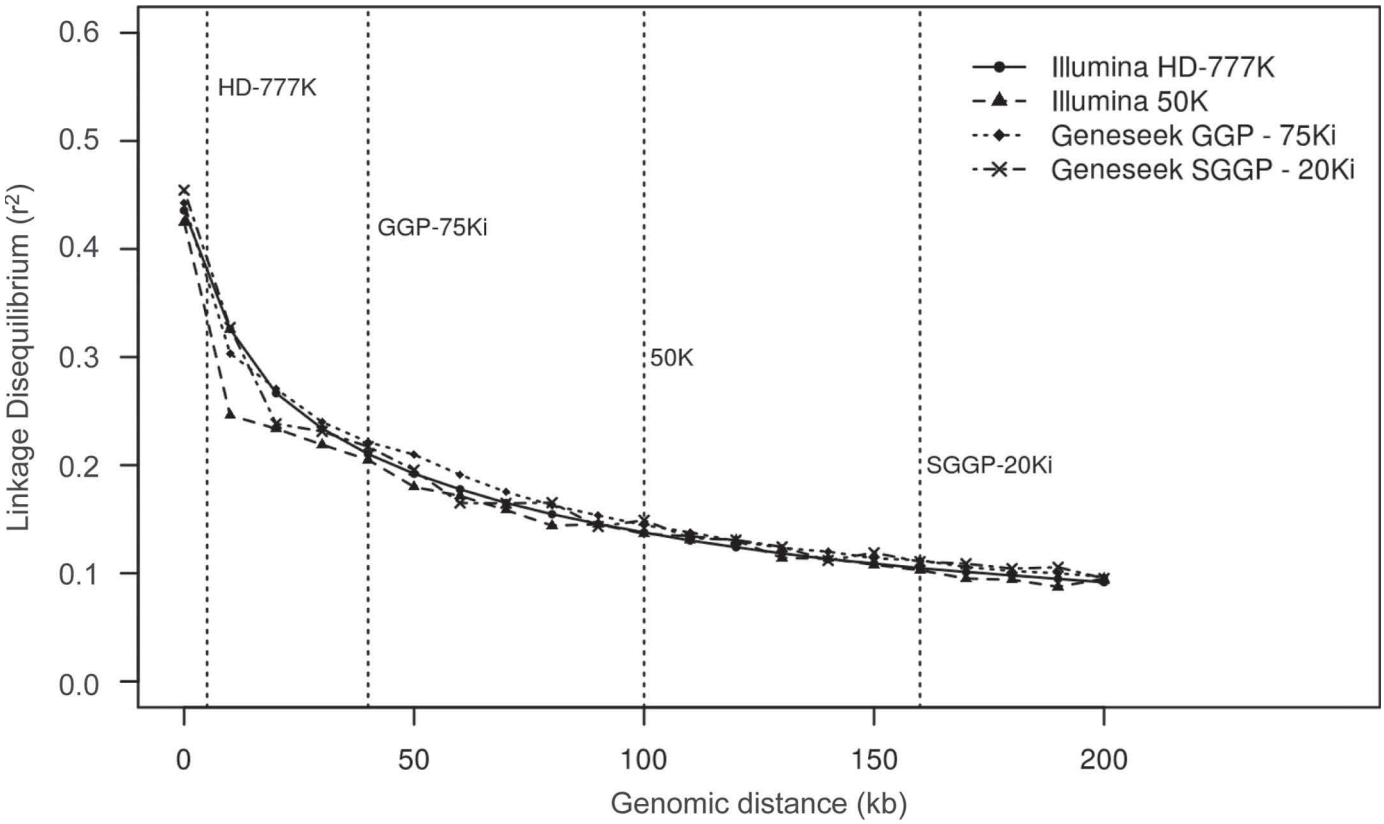

Figure 2. Linkage disequilibrium (r^2) decay over genomic distance in kilobytes; r^2 values are binned at an interval of 10 kb from 0 to 200 kb. Average intermarker distance for Illumina HD-777K (San Diego, CA), GeneSeek GGP-75Ki (Lincoln, NE), Illumina 50K, and GeneSeek SGGP-20Ki is shown with the dashed (- - -) lines.

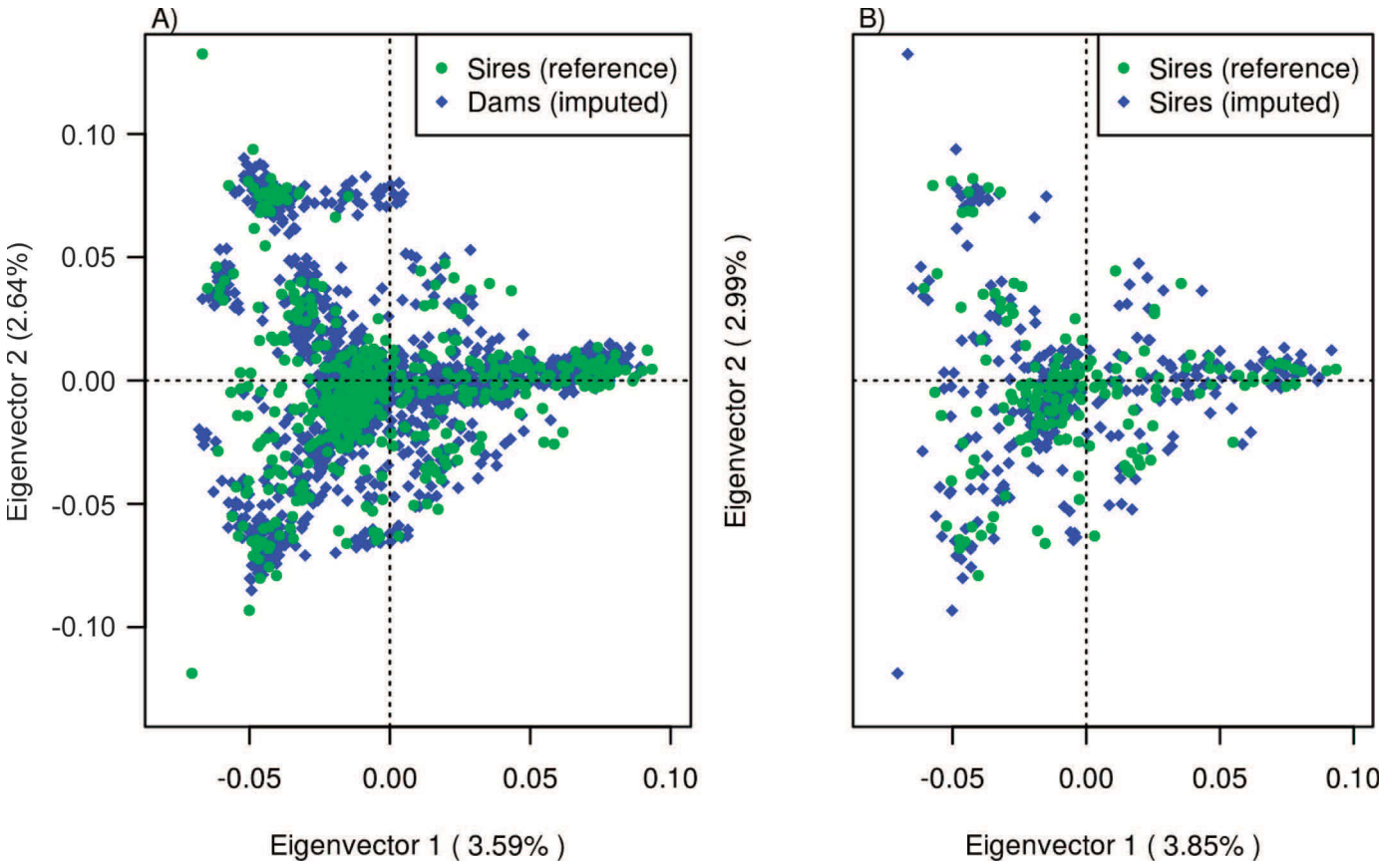

Figure 3. Plots of the first 2 principal components constructed using Illumina 50K (San Diego, CA). (A) Reference animals (all sires, n = 471) and imputed animals (all dams, n = 1,644) used in scenario 1; (B) reference animals (sires, n = 171) and imputed animals (sires, n = 300) used in scenarios 1, 2, 3, and 4. Color version available online.

Table 3. Imputation accuracy (SD) for scenario 1 (imputation from 3K and 7K to 50K) using Beagle v3.3 and v4 and FImpute (Browning and Browning, 2007, 2013; Sargolzaei et al., 2014) with[1] and without pedigree information

			Imputation accuracy[3]	
Chip	Reference/imputed[2]	Software	$corr_{anim}$	ACR_{anim}
LD3K_50K	171/300	FImpute	0.910 (0.052)	0.942 (0.032)
		Beagle v4	0.902 (0.044)	0.935 (0.028)
		Beagle v3.3	0.881 (0.058)	0.924 (0.035)
		FImpute[1]	0.909 (0.033)	0.941 (0.032)
LD7K_50K	171/300	FImpute	0.956 (0.033)	0.971 (0.020)
		Beagle v4	0.943 (0.030)	0.962 (0.019)
		Beagle v3.3	0.934 (0.038)	0.956 (0.024)
		FImpute[1]	0.954 (0.033)	0.970 (0.020)
LD3K_50K	471/1,644	FImpute	0.920 (0.053)	0.948 (0.031)
		Beagle v4	0.918 (0.043)	0.945 (0.027)
		Beagle v3.3	0.903 (0.056)	0.938 (0.033)
		FImpute[1]	0.920 (0.055)	0.948 (0.032)
LD7K_50K	471/1,644	FImpute	0.961 (0.037)	0.974 (0.022)
		Beagle v4	0.955 (0.031)	0.970 (0.019)
		Beagle v3.3	0.949 (0.039)	0.966 (0.023)
		FImpute[1]	0.959 (0.037)	0.973 (0.022)

[1]FImpute was run with pedigree information.

[2]Reference animals were sires (n = 171 or n = 471), and imputed animals were either sires (n = 300) or dams (n = 1,644).

[3]$corr_{anim}$ = correlation between true and imputed genotypes; ACR_{anim} = 1 minus the mean of absolute difference between observed and imputed genotype.

increased imputation accuracy (results not shown). Generally, the difference between software packages diminished with increasing marker density. With marker density of 15K and above as LDP, accuracies were high regardless of the method.

The LDP 3K was too sparse to impute to HD-777K (Figure 4), as accuracies were <0.93. Interestingly, the SGGP-20Ki chip [$corr_{anim}$ = 0.980 (0.018); SD in parentheses] was as accurate as using the 50K [$corr_{anim}$ = 0.982 (0.016)], although the number of markers was about 36.7% greater for the 50K. Although the average accuracies were highest for FImpute, we observed that imputation accuracy was higher with Beagle v4 for individuals (4% of the total number of imputed individuals) that were imputed with lowest accuracy in FImpute (Figure 4).

The 2-step approach (imputation of 3K and 7K to 50K, then to HD-777K) resulted in slightly higher accuracies than using a single step (Supplementary Table S2; http://dx.doi.org/10.3168/jds.2014-9213). The increases were more pronounced for Beagle than for FImpute and also for ACR_{anim} than for $corr_{anim}$.

Imputation of LDP (3K, 7K, SGGP-20Ki, and 50K) to HDP (GGP-75Ki). Imputation accuracy for scenario 3 followed the same trends as observed for scenario 1 and 2. It is worth mentioning that accuracy for the LDP 50K was lower than the LDP SGGP-20Ki (Table 4). This is primarily due to the density of the 50K chip for this scenario. After extracting the 50K markers from the GGP-75Ki, only 7,239 markers remained, whereas the SGGP-20Ki had 13,854 SNP in common with the GGP-75Ki SNP chip. We also observed that results from imputing 3K and 7K to GGP-75Ki were disappointingly low compared with that of imputing 3K and 7K to HDP. Furthermore, even though both the GGP-75Ki and SGGP-20Ki have been selected to contain informative markers for *Bos indicus* breeds, imputation accuracy from SGGP-20Ki to HD-777K [$corr_{anim}$ = 0.980 (0.018); Figure 4] was slightly higher than to GGP-75Ki [$corr_{anim}$ = 0.975 (0.023); Table 4].

Imputation of Ungenotyped Individuals Using FImpute with Pedigree Information

Table 5 shows the imputation accuracy of imputing ungenotyped individuals with 4 or more genotyped offspring using FImpute. Accuracy increased with increasing offspring information. The average imputation accuracy varied more for individuals with less than 6 genotyped offspring (SD of $corr_{anim}$ = 0.024) than individuals with more than 10 genotyped progeny. Additionally, some individuals had 0.05 to 3.9% of their genotypes not inferred or imputed.

Imputation Accuracy with Dosages Instead of Discrete Genotypes

Beagle provides analyses using allele dosages as well as discrete genotypes, and results from other studies have shown an increase in accuracy when dosages are used instead of discrete genotypes. We also observed an increase in accuracy for Beagle when $corr_{anim}$ values were computed with dosages (values range from 0 to 2) instead of discrete genotypes (0, 1, and 2; Table 6). The increase was more pronounced for imputing 3K genotypes to 50K than 7K to 50K. However, the increase was not consistent for the scenario with HD-777K as HDP (results not shown). Despite improvement of imputation accuracies due to using dosages, results of Beagle were still outperformed by FImpute that only yields discrete genotypes except for LDP 3K.

Table 4. Imputation accuracies (SD) for scenario 3 (imputation from 3K, 7K, SGGP-20Ki, to GGP-75Ki) using FImpute without pedigree information and Beagle v4 (Browning and Browning, 2013; Sargolzaei et al., 2014)

			Imputation accuracy[2]	
Chip	Reference/imputed[1]	Software	$corr_{anim}$	ACR_{anim}
LD3K_75Ki	171/300	FImpute	0.845 (0.090)	0.914 (0.032)
		Beagle v4	0.829 (0.079)	0.902 (0.042)
LD7K_75Ki	171/300	FImpute	0.924 (0.058)	0.956 (0.032)
		Beagle v4	0.900 (0.057)	0.940 (0.032)
LD20Ki_75Ki	171/300	FImpute	0.975 (0.023)	0.985 (0.013)
		Beagle v4	0.953 (0.029)	0.972 (0.017)
LD50K_75Ki	171/300	FImpute	0.949 (0.042)	0.970 (0.024)
		Beagle v4	0.923 (0.046)	0.954 (0.027)

[1]Reference (n = 171) and imputed (n = 300) animals were all sires.

[2]$corr_{anim}$ = correlation between true and imputed genotypes; ACR_{anim} = 1 minus the mean of absolute difference between observed and imputed genotype.

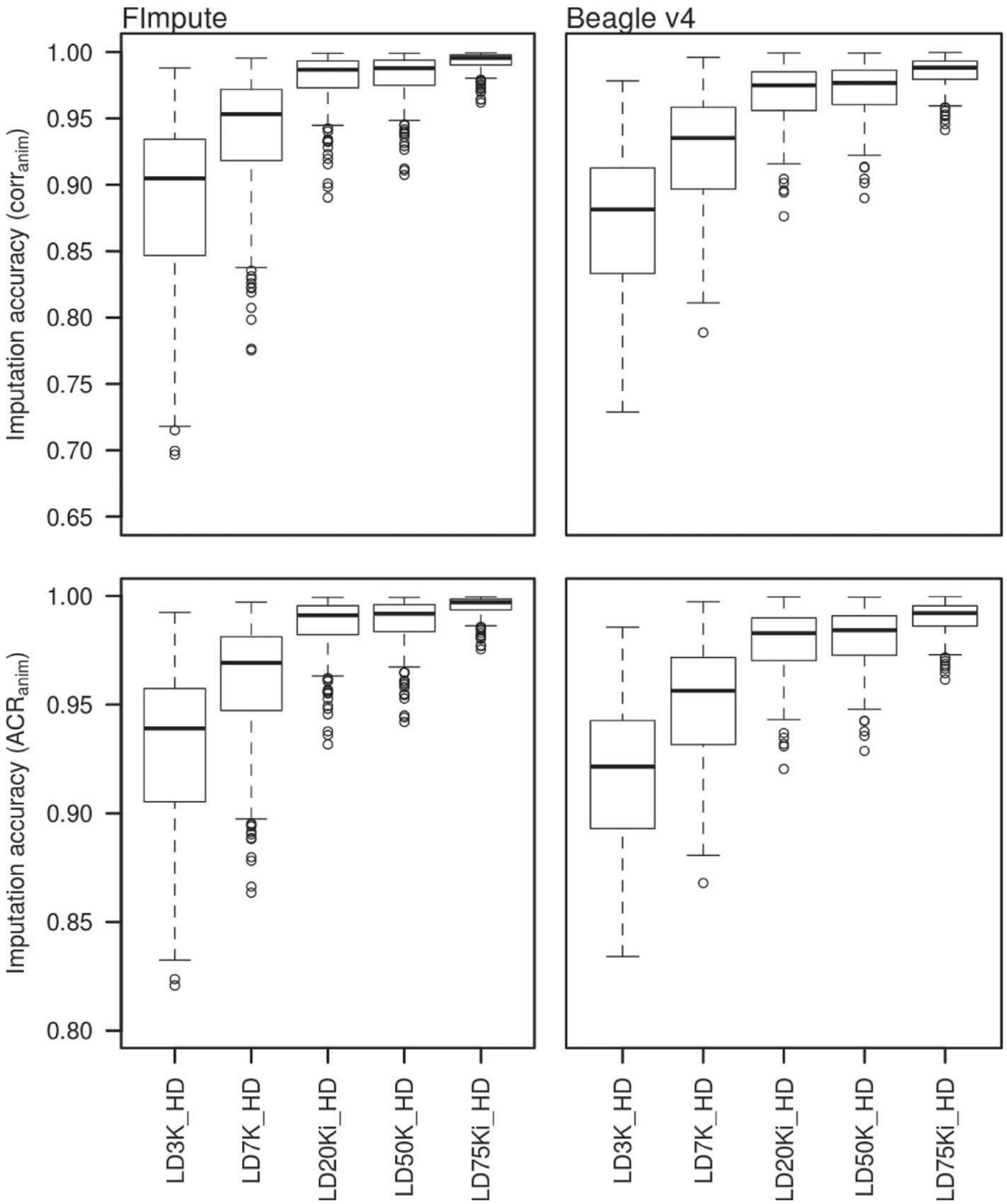

Figure 4. Sample-specific imputation accuracy for lower-density panels (3K, 7K, SGGP-20Ki, 50K, and GGP-75Ki) to high-density panel (HD-777K) using FImpute and Beagle v4 (Browning and Browning, 2013; Sargolzaei et al., 2014). $corr_{anim}$ = correlation between true and imputed genotypes; ACR_{anim} = 1 minus the mean of absolute difference between observed and imputed genotype.

Table 5. Imputation accuracy (SD) for ungenotyped[1] individuals using FImpute (Sargolzaei et al., 2014) with pedigree information

Number of offspring	Number of observations	Imputation accuracy[2]	
		$corr_{anim}$	ACR_{anim}
4–6	38	0.946 (0.024)	0.960 (0.018)
6–10	30	0.974 (0.014)	0.981 (0.010)
11–20	19	0.985 (0.011)	0.990 (0.010)
>20	11	0.998 (0.001)	0.999 (0.001)

[1]Ungenotyped animals were imputed from 0K to the 50K SNP chip.
[2]$corr_{anim}$ = correlation between true and imputed genotypes; ACR_{anim} = 1 minus the mean of absolute difference between observed and imputed genotype.

SNP-Specific Imputation Accuracy and Effect of MAF on Accuracy

For simplicity and clarity, only the HD-777K as the HDP (scenario 2) will be discussed in this section. Similar trends were observed for the other scenarios. Results from FImpute and Beagle v4 are presented in Table 7; those of Beagle v3.3 were lower. The ACR is known to be strongly related to concordance rate (CR). Concordance rate is calculated as the proportion of correctly imputed genotypes. We could approximate CR based on estimates of ACR using $CR = 1 - [x(1 - ACR_{anim})]$, where x was estimated to be 1.89 (SD = 0.03), ranging from 1.84 to 1.99. This shows that imputation from one homozygote to the other is quite rare (Ma et al., 2013).

We also observed that SNP-specific accuracies were sensitive to the amount of reference information available, thus increasing the reference population and marker density reduced error rate for most SNP. However, some SNP still had low accuracies even though marker density and reference information were large. A clear example is a region on chromosome 1, around 44.8 to 45.3 Mb (Supplementary Figure S2; http://dx.doi.org/10.3168/jds.2014-9213).

Based on the results from FImpute, the effect of MAF was opposite between $corr_{snp}$ and ACR_{snp} (Figure 5). The same trend was observed with Beagle v4 and also for the other scenarios with 50K and GGP-75Ki as HDP. However, the effect was observed to be stronger for Beagle v4 than for FImpute. The ACR_{snp} was higher when MAF was low and decreased gradually with increasing MAF, whereas $corr_{snp}$ was lower when MAF was low and increased sharply with increasing MAF. Generally, the effect was seen to be biggest when MAF was <0.05 (Figure 5); this suggests that rare alleles or haplotypes were difficult to impute.

Discarding SNP Due to Predicted Allelic R^2 of Beagle

To increase overall imputation accuracy and reduce the possibility of using poorly imputed SNP for further analysis, Beagle estimates accuracy of imputing a SNP and terms it allelic R^2. We studied the effect of deleting markers based on the predicted allelic R^2. Table 8 shows imputation accuracy for discarding SNP predicted to have lower imputation accuracy from Beagle with the allelic R^2 estimate. Results are not presented for LDP 3K and 7K because even a threshold of about 0.60 caused the deletion of about 75 and 32% of the markers for 3K and 7K, respectively.

Overall imputation accuracy increased slightly by about 0.014, 0.012, and 0.005 correlation points for SGGP-20Ki, 50K, and GGP-75Ki with an allelic R^2 threshold of 0.70. A threshold of 0.70 deleted an average of 6.0% (values ranged between 3 and 8%) of the original number of SNP for LDP markers. We observed that increasing the allelic R^2 value to 0.80 deleted about 23.5% of the SNP for LDP SGGP-20Ki and 50K.

Table 6. Imputation accuracy (SD) using discrete genotypes (0, 1, and 2) versus allele dosages (values ranges from 0 to 2) for lower-density panel 3K and 7K to high-density panel 50K

Chip	Reference/imputed[1]	Software[2]	$corr_{anim}$[3]	
			Genotype	Dosage
LD3K_50K	171/300	Beagle v4	0.902 (0.044)	0.920 (0.032)
		FImpute	0.910 (0.053)	—[4]
LD7K_50K	171/300	Beagle v4	0.943 (0.030)	0.947 (0.023)
		FImpute	0.956 (0.033)	—[4]
LD3K_50K	471/1,644	Beagle v4	0.918 (0.043)	0.926 (0.021)
		FImpute	0.920 (0.054)	—[4]
LD7K_50K	471/1,644	Beagle v4	0.955 (0.031)	0.956 (0.021)
		FImpute	0.961 (0.036)	—[4]

[1]Reference animals were sires (n = 171 or n = 471), and imputed animals were either sires (n = 300) or dams (n = 1,644).
[2]Browning and Browning, 2013; Sargolzaei et al., 2014.
[3]$corr_{anim}$ = correlation between true and imputed genotypes.
[4]FImpute does not give dosage rate.

Table 7. Summary statistics of SNP-specific imputation accuracy (SD) of lower-density panels (3K, 7K, 50K, and SGGP-20Ki, GGP-75Ki) to higher-density panel HD-777K with FImpute without pedigree information and Beagle v4 (Browning and Browning, 2013; Sargolzaei et al., 2014)

			Imputation accuracy[2]	
Chip	Reference/imputed[1]	Software	$corr_{snp}$	ACR_{snp}
LD3K_HD	171/300	FImpute	0.735 (0.140)	0.929 (0.047)
		Beagle v4	0.683 (0.173)	0.917 (0.055)
LD7K_HD	171/300	FImpute	0.859 (0.098)	0.962 (0.029)
		Beagle v4	0.817 (0.125)	0.950 (0.036)
LD20Ki_HD	171/300	FImpute	0.950 (0.061)	0.987 (0.015)
		Beagle v4	0.918 (0.082)	0.979 (0.020)
LD50K_HD	171/300	FImpute	0.957 (0.059)	0.988 (0.015)
		Beagle v4	0.928 (0.078)	0.981 (0.020)
LD75Ki_HD	171/300	FImpute	0.981 (0.042)	0.996 (0.010)
		Beagle v4	0.957 (0.063)	0.990 (0.013)

[1]Reference animals (sires, n = 171) and imputed animals (sires, n = 300).

[2]$corr_{snp}$ = correlation between true and imputed genotypes; ACR_{snp} = 1 minus the mean of absolute difference between observed and imputed genotype.

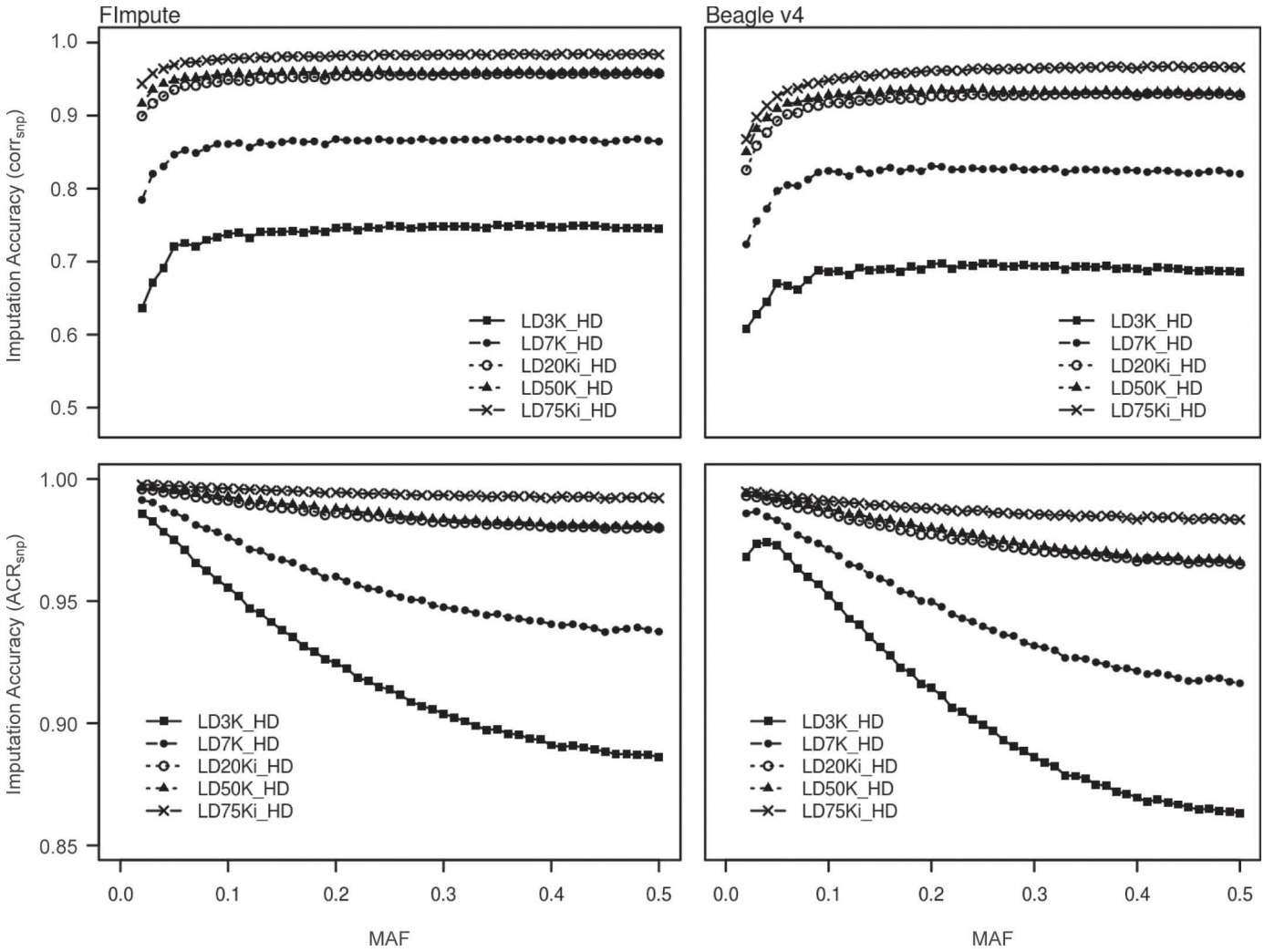

Figure 5. Imputation accuracy against minor allele frequency (MAF) for lower-density SNP panels (3K, 7K, SGGP-20Ki, 50K, and GGP-75Ki) to high-density panel (HD-777K) using FImpute and Beagle v4 (Browning and Browning, 2013; Sargolzaei et al., 2014); MAF was binned at 0.01 intervals. $corr_{snp}$ = correlation between true and imputed genotypes; ACR_{snp} = 1 minus the mean of absolute difference between observed and imputed genotype.

Table 8. Proportion of SNP markers and imputation accuracies (SD) after removing markers with predicted allelic R^2 below threshold of 0.70 to 0.90 with Beagle v4 (Browning and Browning, 2013)

Chip	Threshold	Imputation accuracy[1]		Proportion. of markers excluded[2]
		$corr_{snp}$	ACR_{snp}	
LD20Ki_HD	0.00	0.918 (0.082)	0.979 (0.020)	0.00
	0.70	0.932 (0.051)	0.980 (0.017)	0.08
	0.80	0.944 (0.043)	0.932 (0.014)	0.25
	0.85	0.955 (0.038)	0.987 (0.011)	0.45
	0.90	0.968 (0.035)	0.992 (0.008)	0.68
LD50K_HD	0.00	0.928 (0.078)	0.981 (0.020)	0.00
	0.70	0.940 (0.050)	0.982 (0.017)	0.06
	0.80	0.951 (0.028)	0.985 (0.014)	0.22
	0.85	0.962 (0.037)	0.989 (0.011)	0.39
	0.90	0.973 (0.033)	0.994 (0.008)	0.60
LD75Ki_HD	0.00	0.957 (0.063)	0.990 (0.013)	0.00
	0.70	0.962 (0.045)	0.991 (0.011)	0.03
	0.80	0.967 (0.038)	0.991 (0.010)	0.09
	0.85	0.972 (0.034)	0.993 (0.009)	0.19
	0.90	0.980 (0.030)	0.995 (0.007)	0.38

[1]$corr_{snp}$ = correlation between true and imputed genotypes; ACR_{snp} = 1 minus the mean of absolute difference between observed and imputed genotype.

[2]Proportion of markers excluded before computing imputation accuracies. Correlation between $corr_{snp}$ and allelic R^2 was 0.812, 0.818, and 0.758 for SGGP-20Ki, 50K, and GGP-75Ki, respectively.

Generally, the correlation between allelic R^2 and $corr_{snp}$ was high and ranged from 0.78 to 0.86 depending on the density of the LDP. Whereas allelic R^2 present a possibility of reducing imputation error rate, the approach also presents a minor challenge, as markers with high $corr_{snp}$ could be deleted due to low predicted allelic R^2 (Figure 6).

Effect of Relatedness on Imputation Accuracy

Animals with a close relative genotyped (e.g., parent, siblings, and other close relatives) should intuitively have higher imputation accuracy than animals with distant or no close relatives available in the reference population. We investigated this by regressing imputation accuracy on varying levels of relatedness (relmax, rel5, rel10, and overall$^+$). We observed a strong and highly significant relationship between relatedness and imputation (Figure 7) for all the relatedness measures. The relationship was stronger using rel5 than the other measures (relmax, rel10, and overall$^+$). Imputation accuracy increased as relatedness increased. In fact, almost all individuals with mean top 5 kinship estimates lower than 0.10 showed imputation accuracies lower than 0.90 regardless of the marker density for the LDP. The sire-to-sire scenario had only a few animals with the mean top 5 relatedness lower than 0.10. The adjusted R^2 for the sire-to-dam scenario of the fitted quadratic model was between 59.4 to 65.8%. However, this was marker density-dependent, with adjusted R^2 decreasing with increasing marker density for the LDP.

Computing Time

Results of computation times of the largest chromosome (BTA 1) for 2 selected scenarios (HDP of 50K and HD-777K) are presented in Figure 8. The 2 scenarios provide a comprehensive overview of the computation efficiency of FImpute and Beagle v3.3 and v4. FImpute was very fast (Figure 8) and more memory efficient (results not shown) than Beagle v4 and v3.3. For a data set of 171 reference individuals and 300 imputed individuals from LDP 3K to HD-777K, computing time was below 60 s. Similar computing times (<60 s) were recorded for the sire (reference, n = 471)-to-dam (imputed, n = 1,644) imputation scenario of 3K and 7K LDP to 50K. Beagle v4 was about 10 to 20 times faster than Beagle v3.3 (Figure 8). To our knowledge, this is the first attempt of comparing the 2 Beagle packages, and the superior result of imputation accuracy and speed of computation for Beagle v4 provide evidence of its advantage over Beagle v3.3, although both versions performed similarly in terms of memory usage (data not shown).

DISCUSSION

We studied the accuracy of SNP imputation for a *Bos indicus* dairy cattle breed (Gyr) using different commercially available SNP chips as low- or high-density marker panels to ascertain their usefulness in maximizing accuracy and possibly reducing genotyping costs. Imputation was undertaken with a deterministic software, FImpute (Sargolzaei et al., 2014, 2008), and

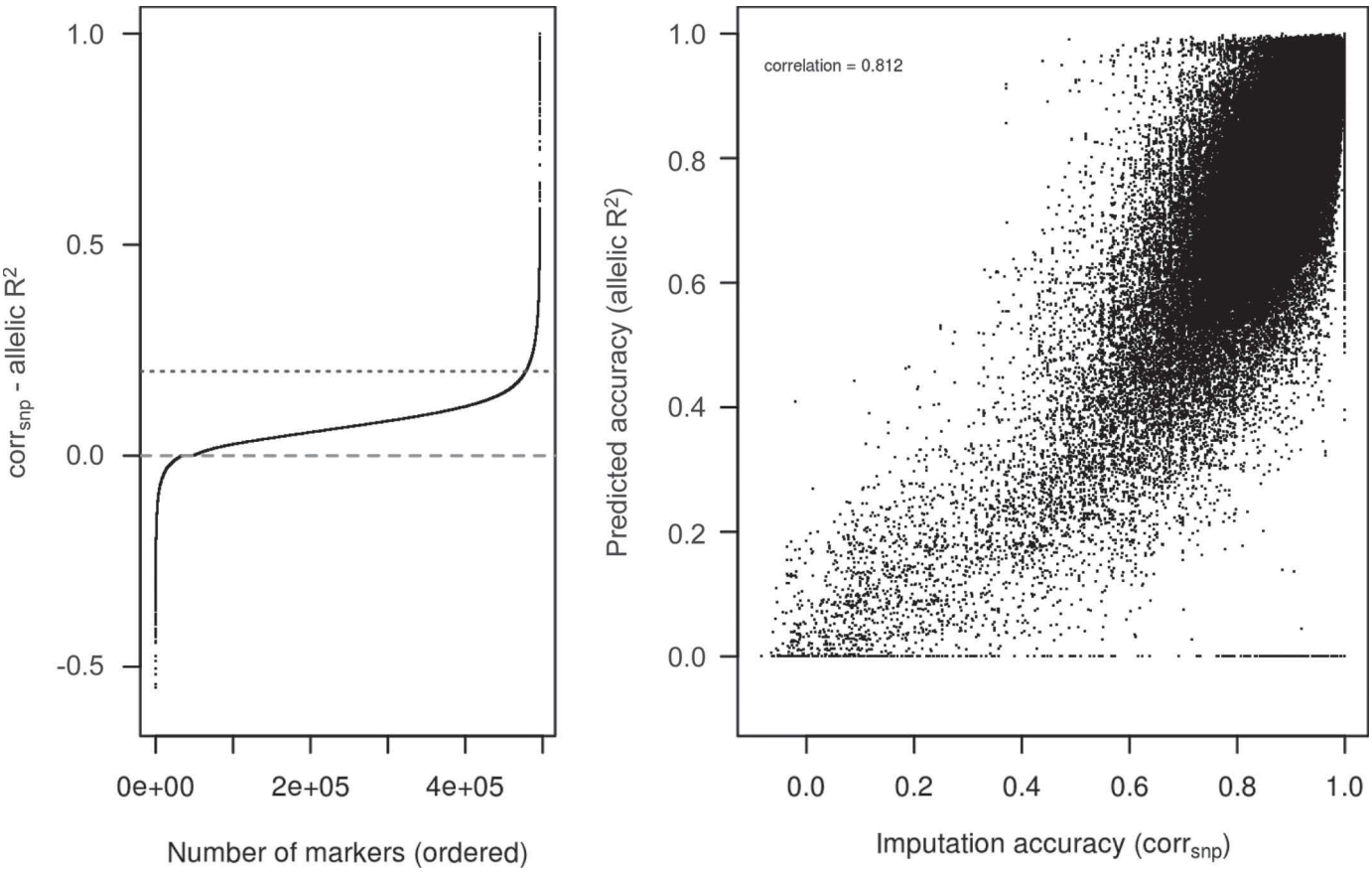

Figure 6. Difference between $corr_{snp}$ and allelic R^2 and plot of $corr_{snp}$ against allelic R^2 for imputation of GeneSeek SGGP-20Ki (Lincoln, NE) to Illumina HD-777K (San Diego, CA). $corr_{snp}$ = correlation between true and imputed genotypes; allelic R^2 = prediction accuracy of Beagle v4 (Browning and Browning, 2013).

a stochastic software, Beagle (Browning and Browning, 2013, 2007), both old (v3.3) and new (v4) versions.

Marker Density, Allele Frequency Distribution, and LD

The SGGP-20Ki and GGP-75Ki both retained about 85 to 90% of their original number of markers, unlike the 50K, which retained about 52%, based on results from the current study as well as from the study of Carvalheiro et al., (2014). The MAF trends of the 50K and HD-777K chip, as shown in Figure 1 for Gyr (*Bos indicus*), is markedly different from what has been observed in *Bos taurus* breeds (Wiggans et al., 2012, 2013; Pérez O'Brien et al., 2014). The high level of monomorphic markers especially for the 50K might be problematic for several reasons. There is an indication that imputation accuracy is affected when imputing to HD-777K. This is deduced from the fact that, although it had about 35% more SNP than the SGGP-20Ki, accuracies were very similar. An even higher accuracy for the SGGP-20Ki chip was observed for the Nelore breed (Carvalheiro et al., 2014). In addition, having about 35% of the markers with MAF between 0.02 and 0.10 would require larger data sets to be able to map QTL (Ardlie et al., 2002), as well as to generate reasonable prediction accuracies in GS when the 50K chip is used.

Linkage disequilibrium estimates with 50K and HD-777K were similar to what has been reported (Pérez O'Brien et al., 2014). The results are consistent for this breed and other *Bos indicus* breeds (Neves et al., 2014; Pérez O'Brien et al., 2014; Porto-Neto et al., 2014). Conversely, r^2 was consistently lower than what has been reported for taurine breeds (Larmer et al., 2014; Pérez O'Brien et al., 2014; Porto-Neto et al., 2014). Average r^2 >0.10 for SGGP-20Ki, 50K, and GGP-75Ki seems to be reasonable to achieve high imputation accuracy. However, this conclusion can be confounded with the effect of marker density and number of genotyped reference individuals. Conversely, the lower LD levels for the 50K and SGGP-20Ki might not be enough to effectively capture QTL effect for genomic prediction

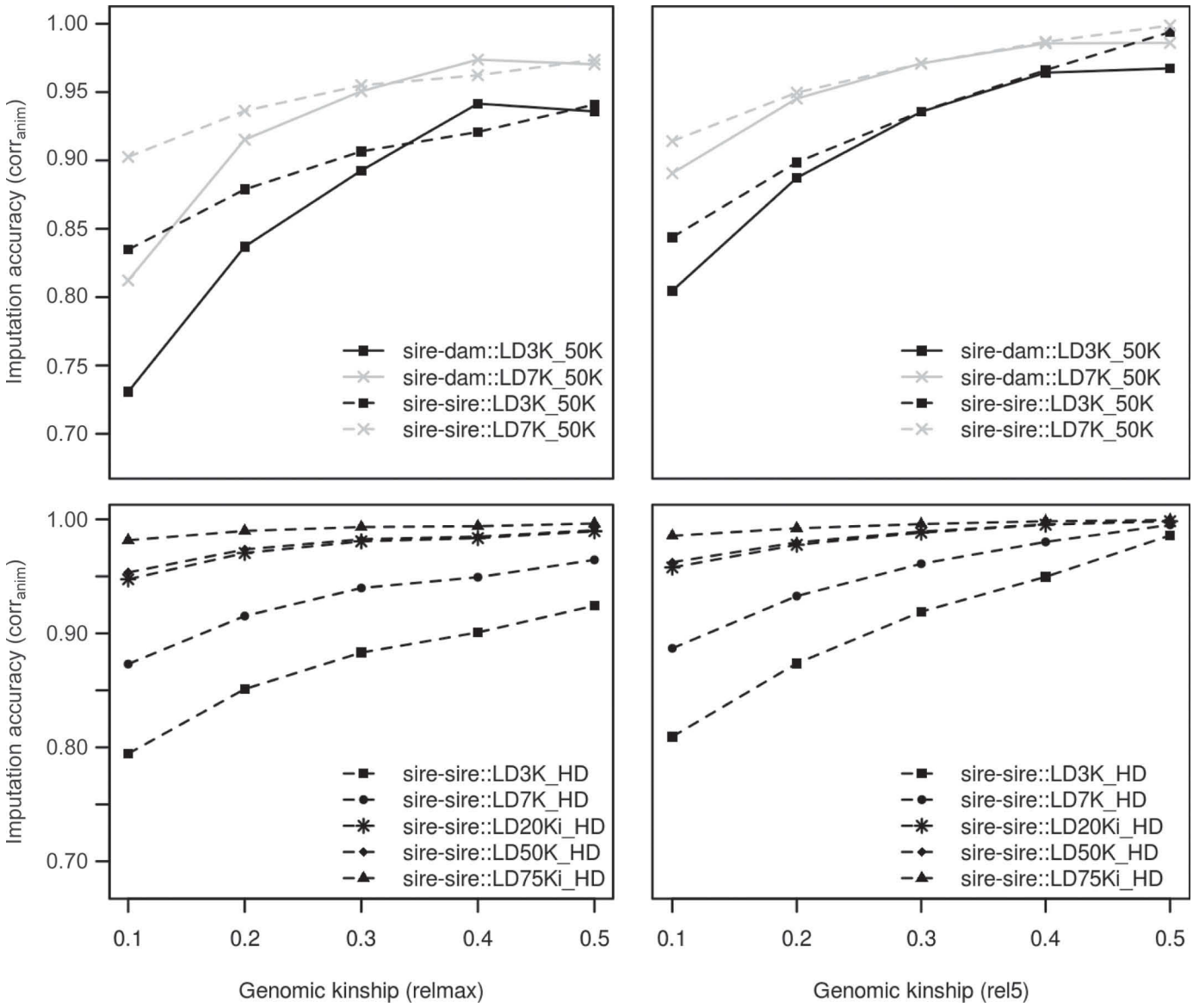

Figure 7. Plot of imputation accuracy as a function of relatedness to the reference set. Maximum (relmax) and average of top 5 (rel5) relationships of imputed individual to the reference set. $corr_{anim}$ = correlation between true and imputed genotypes.

(Calus et al., 2008) and genome-wide association study (Ardlie et al., 2002). We concluded that constraints exist in using 50K as HDP for *Bos indicus* breeds, whereas the level of LD of syntenic markers for the GGP-75Ki and HD-777K seems reasonably high for use as HDP in the implementation of GS.

Imputation Accuracy with Different SNP Chips

Several studies have evaluated imputation accuracy for *Bos taurus* beef and dairy cattle breeds (Brøndum et al., 2012; Khatkar et al., 2012; Hozé et al., 2013; Ma et al., 2013; Pryce et al., 2014; Sargolzaei et al., 2014; Schrooten et al., 2014). However, to our knowledge, only a few imputation studies have been undertaken and reported in indicine cattle breeds (Bolormaa et al., 2013; Carvalheiro et al., 2014). In our study, imputation accuracy was within the range of those reported for *Bos taurus* breeds (Weigel et al., 2010b; Dassonneville et al., 2011; Mulder et al., 2012; Ma et al., 2013; Wiggans et al., 2013; Berry et al., 2014) and *Bos indicus* breeds (Bolormaa et al., 2013; Carvalheiro et al., 2014). It is difficult to compare results between populations and between *Bos indicus* and *Bos taurus* breeds. Differences

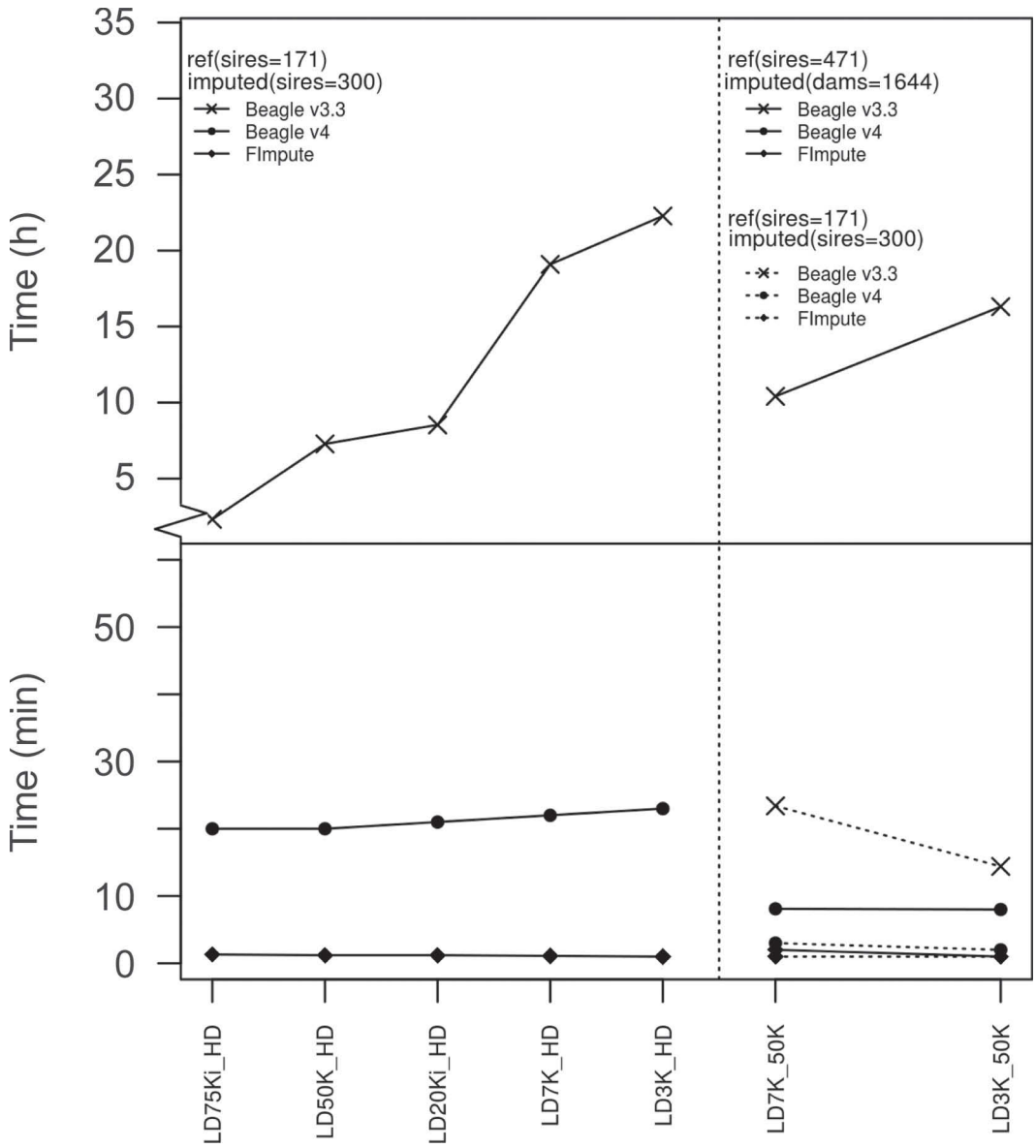

Figure 8. Computing time for chromosome 1 of FImpute and Beagle v4 and v3.3 (left of the dashed line; Browning and Browning, 2007, 2013; Sargolzaei et al., 2014); computing time for imputation of 3K, 7K, SGGP-20Ki, 50K, and GGP-75Ki to HD-777K with 171 reference animals and 300 imputed set using FImpute and Beagle v4 and v3.3 (right of the dashed line); computing time for imputation of 3K and 7K to 50K with 171 reference animals and 300 imputed set or 471 reference animals and 1,644 imputed set using FImpute, Beagle v4 and v3.3.

in population structure (e.g., LD; also confounded with marker density of LDP), size of reference population (Berry et al., 2014), and relationship between reference and imputed population all hamper comparisons across studies.

Imputation from LDP 3K and 7K to HDP 50K and GGP-75Ki were lower than using SGGP-20Ki, 50K and 75K to HD-777K. In addition, imputation of 3K and 7K to HD-777K was also lower than imputing to 50K. These results have also been observed in several other studies (Ma et al., 2013; Berry et al., 2014), but results from Sargolzaei et al. (2014) indicate that, imputation from 3K and 7K are highest when genotype information is available on parent and or grandparent of the LDP individuals. Imputation accuracy for the LDP 3K and 7K also increased when 2-step imputation approach (3K and 7K to 50K then to HD-777K) was used. VanRaden et al. (2013) and Larmer et al. (2014) also reported increases in accuracy using a 2-step approach for LDP 3K and 7K.

Interestingly, imputation of LDP SGGP-20Ki to HDP GGP-75Ki and HD-777K was very accurate, and SGGP-20Ki may be recommended as the low-density SNP panel of choice for this population. The result obtained in Nelore (Carvalheiro et al., 2014) and in Gyr in the current study suggests that this conclusion could be valid for other *Bos indicus* breeds. Imputation of 50K to HD-777K was even more accurate than to GGP-75Ki. This is directly related to the number of markers the 50K and GGP-75Ki share. The low number of common markers of these 2 SNP chips presents a challenge for breeding organization and researchers that had initially genotyped a substantial number of animals on 50K. The best, and somehow cheaper, approach when a large number of individuals have already been genotyped on the 50K chip would be to adopt the HD-777K as the HDP. Strategies suggested by Druet et al. (2014) and Yu et al. (2014) could be used to select individuals to be genotyped on HD-777K with the aim of maximizing imputation accuracy. However, we recommend that subsequent genotyping should be undertaken with the SGGP-20Ki chip to reduce cost. Moreover, the price of genotyping SGGP-20Ki is currently about 25 to 45% lower than that of the 50K and could be considered as an LDP for *Bos indicus* populations.

Increasing the proportion of reference animals to ~2:1 caused a slight increase in accuracy. The increase could be directly attributed to the increase in the number of rare haplotypes and increase of related individuals added to the reference population. Additionally, the increase in reference population also increased the relatedness between the reference and imputed set and thus accuracies improved. This trend has also been reported in several other studies (Khatkar et al., 2012; Badke et al., 2013; Pausch et al., 2013).

FImpute without pedigree information was the superior software package in our study. However, using pedigree information in FImpute gave similar and sometimes slightly lower accuracies than scenarios without prior pedigree information. This can be attributed to pedigree (Sargolzaei et al., 2014) and genotyping errors, although a Mendelian error threshold was set. Sargolzaei et al. (2014) argues that, with a large reference population and relatively large low density marker panel (e.g., 50K), imputation would be accurate even if family information is ignored.

Imputing ungenotyped individuals with more than 4 genotyped offspring was very accurate in this population. FImpute uses pedigree information to identify ungenotyped individuals with more than 4 (minimum allowed by the software) genotyped progeny to impute ungenotyped sires or dams. Initial filling of missing genotypes is done with offspring information. Subsequently, information from the population is used if present. In our study, population information from more than 1,500 genotyped animals was available. The accuracies reported here are similar or slightly higher compared with the results of Nicolazzi et al. (2013), Bouwman et al. (2014), and Boison et al. (2014). The reason for the difference could be attributed to the higher level of relatedness in this population. For imputation of ungenotyped individuals to be part of routine imputation process for most breeds with no DNA samples of important animals (founders, accurate phenotypic information, and new phenotypes) in the pedigree, verification of offspring information should be undertaken to reduce errors. The approach of Calus et al. (2011) in finding sib errors could be used. Calus et al. (2011) used the principle of opposing homozygous genotypes based on Mendelian inconsistencies to identify full- and half-sib errors. We recommend this be done before offspring information is used to impute ungenotyped sires or dams.

Imputation accuracy for SNP has been reported to be dependent on the measure of accuracy used (ACR_{snp} and $corr_{snp,}$). Generally, accuracy is higher for markers with low MAF and decreases with increasing MAF when ACR_{snp} is used (Hickey et al., 2012a; Badke et al., 2013; Ma et al., 2013; Brøndum et al., 2014; Sargolzaei et al., 2014). With $corr_{snp}$, accuracy is lowest for markers with low MAF and increases with increasing MAF (Hickey et al., 2012a; Badke et al., 2013; Ma et al., 2013; Brøndum et al., 2014; Sargolzaei et al., 2014). Hickey et al. (2012a), Ma et al. (2013), and Badke et al. (2013) have demonstrated the importance of using $corr_{snp}$ instead of ACR_{snp} in measuring SNP-specific accuracy.

Poor imputation accuracies for certain regions of the genome were observed, and other studies (Erbe et al., 2012; Pausch et al., 2013; Berry et al., 2014; Carvalheiro et al., 2014) have also reported this in several populations. Generally, the ends of chromosomes are poorly imputed and this is due to the limited number of tag SNP. However, other regions, even with sufficient amount of neighboring markers have continuously been imputed poorly. Several reasons have been cited (Erbe et al., 2012; Pausch et al., 2013; Berry et al., 2014; Carvalheiro et al., 2014), with wrong assembly of such regions being the most prominent reason (Pausch et al., 2013; Carvalheiro et al., 2014). A clear example was observed on BTA 1 between 44.8 and 45.3 Mb. Carvalheiro et al. (2014) reported low LD between the surrounding markers of this region. In addition, visual checks of the ENSEMBL biomart platform (Flicek et al., 2014) reveal that the area is spanned with many

small contigs, instead of a single contig for such a small region.

Excluding SNP Based on Predicted Accuracy (Allelic R^2)

Dealing with SNP that are imputed with poor accuracy is an important issue as it reduces the potential effect when the genotypes are used in subsequent studies (e.g., genome-wide association study, GS). Generally, the use of allele dosages (values ranging from 0 to 2) instead of the most probable genotypes have been deemed to be a way to reduce the overall effect of poorly imputed SNP (Marchini and Howie, 2010; Calus et al., 2014). Dosages have been recommended because they tend to reduce bias especially in genomic prediction and association studies (Marchini and Howie, 2010; Calus et al., 2014). However, the use of dosages might not be satisfactory. The ability to discard SNP that have been imputed with poor accuracy has become much more important with the imputation of LDP to full sequence data. For example, in cattle populations, imputation of LDP (e.g., 50K) to HD-777K had largely been undertaken with large reference population (n = >1,000), thus accuracies were extremely high (>0.97). However, imputation to full sequence data was undertaken with few individuals (mostly ancestors or individuals with largest contribution of genes to the current population) that have carefully been selected to maximize imputation accuracy. This lowered the accuracy of imputing rare haplotypes and the overall imputation accuracy has been lower than those of HD-777K or 50K (Bouwman and Veerkamp, 2014; Brøndum et al., 2014; van Binsbergen et al., 2014).

The predicted allelic R^2 by Beagle could be used to further discard SNP that have been imputed poorly; in our study it gave a good prediction of the true imputation accuracy of an SNP. The correlation between allelic R^2 and $corr_{snp}$ was between 0.78 and 0.86 depending on the LDP. van Binsbergen et al. (2014) also reported similar correlation between allelic R^2 and $corr_{snp}$ of 0.79 for HD-777K to full sequence imputation. However, Figure 6 points at the bias in the estimations, especially the over estimation of true accuracy by allelic R^2 (van Binsbergen et al., 2014). As with $corr_{snp}$, allelic R^2 was also influenced by the size of the reference panel, low-frequency alleles, as well as marker density of LDP. The results of the current study suggest that an allelic R^2 threshold of about 0.70 (~20K markers imputed to HD) or 0.80 (~65K markers imputed to HD-777K), which deleted on average 8% of the imputed markers, is a reasonable threshold to adopt. However, in the case of imputing 50K or HD-777K to full sequence, a threshold that will not delete too many imputed markers needs to be found.

Effect of Relatedness on Accuracy of Imputation

To assist in the comparison of imputation accuracy across population, we recommend that summary of relatedness between the reference and imputed individuals should be routinely reported. Relatedness measures such as traceability (pedigree-based; Zhang and Druet, 2010; Mulder et al., 2012) or average maximum and top 5 and 10 genomic kinships between reference and imputed individuals could be reported. Daetwyler et al. (2013) suggested that scientific literature on genomic evaluations should also contain relatedness between the reference and validation individuals so as to better contextualize the results.

In the current study, we observed a strong relationship between rel5 and imputation accuracy. Other kinship estimates, such as relmax, rel10, and the overall$^+$ positive kinship estimates between reference and imputed sets, were also a good predictor of imputation accuracy. However, the adjusted R^2 was largest and ranged between 36.5 and 68% using rel5.

The estimate of rel5 could be used in a strategy (a) to exclude individuals that would be imputed poorly and (b) for sequential imputation of individuals' genotypes on the LDP. Sequential imputation means that a subset of animals with highest rel5 could be imputed first and then subsequently added to the reference panel for the imputation of the next related batch. In a simple demonstration, consider a case where 2 generation (e.g., sire and offspring) of a family is present for imputation. The grand sire has been genotyped on the HDP panel and both sire and offspring are genotyped on a LDP. It will be best when the sire is imputed first and added to the reference population before the offspring is imputed (Figure 7). Instead of using traceability, which is pedigree-based, rel5 could be used.

Computational Efficiency

A large computational advantage (speed and memory usage) was observed for FImpute, and this has also been reported elsewhere (Ma et al., 2013; Larmer et al., 2014; Sargolzaei et al., 2014). It has thus been suggested as an attractive method for use in livestock populations with a need of routine imputation of individuals for genomic prediction.

Beagle v4 was recently released by Browning and Browning (2013) with improvement in the accuracy of genotype calling, phasing, imputation, and identity-by-descent detection. Based on the current study, not

only did imputation accuracy increase compared with Beagle v3.3, the speed at which genotypes of reference individuals are phased and LDP individuals imputed has also been improved. FImpute and Beagle v4 would become increasingly more important especially in this sequence imputation era due to the amount of time that would be needed to undertake imputation for large numbers of LDP individuals to full sequence genotypes.

CONCLUSIONS

Genotype imputation in Gyr was accurate across several SNP chips. Linkage disequilibrium estimates, marker density, and its distribution across MAF intervals with the Illumina 50K were not satisfactory as an HDP panel for this population. Both the GeneSeek GGP-75Ki and Illumina HD-777K could be a reasonable choice. We concluded that using LDP with more than 15K markers to impute HD-777K was very accurate. In that regard, GeneSeek SGGP-20Ki gave very similar or higher imputation accuracies compared with 50K when imputed to HD-777K or GGP-75Ki, respectively. It is important to note that, when GGP-75Ki was considered as HDP, imputation accuracy for 50K was observed to be poor, due to the fact that both chips share less than 10K markers. We also concluded that rel5 is a robust predictor of imputation accuracy accounting for about 36.5 to 68% of the variation observed in sample-specific accuracies. It could be adopted and routinely reported in imputation studies to facilitate further comparison between populations. Beagle v4 and FImpute gave more accurate results than Beagle 3.3. Beagle v4 was computationally advantageous to Beagle v3.3 and should be the method of choice. Finally, the potential effect of using any of the SNP chips and imputed genotypes on accuracy of genomic predictions should be comprehensively evaluated before a final decision on which chip to use as LDP or HDP could be made.

ACKNOWLEDGMENTS

Marcos V. B. Silva was supported by the Embrapa (Brazil) SEG 02.09.07.008.00.00 "Genomic Selection in Dairy Cattle in Brazil," CNPq PVE 407246/2013-4 "Genomic Selection in Dairy Gyr and Girolando Breeds," and FAPEMIG CVZ PPM 00395/14 "Genomic Selection in Brazilian Dairy Breeds" appropriated projects.

REFERENCES

Ardlie, K. G., L. Kruglyak, and M. Seielstad. 2002. Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. 3:299–309. http://dx.doi.org/10.1038/nrg777.

Badke, Y. M., R. O. Bates, C. W. Ernst, J. Fix, and J. P. Steibel. 2014. Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3 (Bethesda) 4:623–631. http://dx.doi.org/10.1534/g3.114.010504.

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. http://dx.doi.org/10.1186/1471-2156-14-8.

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. http://dx.doi.org/10.1017/S1751731111000309.

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

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

Bolormaa, S., J. E. Pryce, K. Kemper, K. Savin, B. J. Hayes, W. Barendse, Y. Zhang, C. M. Reich, B. A. Mason, R. J. Bunch, B. E. Harrison, A. Reverter, R. M. Herd, B. Tier, H.-U. Graser, and M. E. Goddard. 2013. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in *Bos taurus*, *Bos indicus*, and composite beef cattle. J. Anim. Sci. 91:3088–3104. http://dx.doi.org/10.2527/jas.2012-5827.

Bouwman, A. C., J. M. Hickey, M. P. Calus, and R. F. Veerkamp. 2014. Imputation of non-genotyped individuals based on genotyped relatives: Assessing the imputation accuracy of a real case scenario in dairy cattle. Genet. Sel. Evol. 46:6. http://dx.doi.org/10.1186/1297-9686-46-6.

Bouwman, A. C., and R. F. Veerkamp. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genet. 15:105. http://dx.doi.org/10.1186/s12863-014-0105-8.

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. http://dx.doi.org/10.1186/1471-2164-15-728.

Brøndum, R. F., P. Ma, M. S. Lund, and G. Su. 2012. Short communication: Genotype imputation within and across Nordic cattle breeds. J. Dairy Sci. 95:6795–6800. http://dx.doi.org/10.3168/jds.2012-5585.

Brøndum, R. F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. J. Dairy Sci. 94:4700–4707. http://dx.doi.org/10.3168/jds.2010-3765.

Browning, B. L., and S. R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194:459–471. http://dx.doi.org/10.1534/genetics.113.150029.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097. http://dx.doi.org/10.1086/521987.

Calus, M. P., A. de Roos, and R. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561. http://dx.doi.org/10.1534/genetics.107.080838.

Calus, M. P. L., A. C. Bouwman, J. M. Hickey, R. F. Veerkamp, and H. A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal 8:1743–1753. http://dx.doi.org/10.1017/S1751731114001803.

Calus, M. P. L. M., H. A. Mulder, and J. W. M. Bastiaansen. 2011. Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. Genet. Sel. Evol. 43:34. http://dx.doi.org/10.1186/1297-9686-43-34.

Carvalheiro, R., S. A. Boison, H. H. R. Neves, M. Sargolzaei, F. S. Schenkel, Y. T. Utsunomiya, A. M. P. O'Brien, J. Sölkner, J. C. McEwan, C. P. Van Tassell, T. S. Sonstegard, and J. F. Garcia.

2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46:69. http://dx.doi.org/10.1186/s12711-014-0069-1.
Cleveland, M. A., and J. M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci. 91:3583–3592. http://dx.doi.org/10.2527/jas.2013-6270.
Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de Los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365. http://dx.doi.org/10.1534/genetics.112.147983.
Daetwyler, H. D., K. E. Kemper, J. H. J. van der Werf, and B. J. Hayes. 2012a. Components of the accuracy of genomic prediction in a multi-breed sheep population. J. Anim. Sci. 90:3375–3384. http://dx.doi.org/10.2527/jas.2011-4557.
Daetwyler, H. D., A. A. Swan, J. H. J. van der Werf, and B. J. Hayes. 2012b. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. Genet. Sel. Evol. 44:33. http://dx.doi.org/10.1186/1297-9686-44-33.
Dassonneville, R., R. F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M. S. Lund, V. Ducrocq, and G. Su. 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J. Dairy Sci. 94:3679–3686. http://dx.doi.org/10.3168/jds.2011-4299.
Dassonneville, R., S. Fritz, V. Ducrocq, and D. Boichard. 2012. Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. J. Dairy Sci. 95:4136–4140. http://dx.doi.org/10.3168/jds.2011-5133.
Druet, T., and M. Georges. 2010. A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184:789–798. http://dx.doi.org/10.1534/genetics.109.108431.
Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb) 112:39–47. http://dx.doi.org/10.1038/hdy.2013.13.
Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114–4129. http://dx.doi.org/10.3168/jds.2011-5019.
Flicek, P., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino, and S. M. Searle. 2014. Ensembl 2014. Nucleic Acids Res. 42:D749–D755. http://dx.doi.org/10.1093/nar/gkt1196.
Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93:1243–1252. http://dx.doi.org/10.3168/jds.2009-2619.
Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012a. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654. http://dx.doi.org/10.2135/cropsci2011.07.0358.
Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. J. van der Werf, and M. A. Cleveland. 2012b. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44:9. http://dx.doi.org/10.1186/1297-9686-44-9.
Howie, B. N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5:e1000529. http://dx.doi.org/10.1371/journal.pgen.1000529.
Hozé, C., M.-N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. 45:33. http://dx.doi.org/10.1186/1297-9686-45-33.
Huang, Y., J. M. Hickey, M. A. Cleveland, and C. Maltecca. 2012. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet. Sel. Evol. 44:25. http://dx.doi.org/10.1186/1297-9686-44-25.
Hutchison, J. L., J. B. Cole, and D. M. Bickhart. 2014. Short communication: Use of young bulls in the United States. J. Dairy Sci. 97:3213–3220. http://dx.doi.org/10.3168/jds.2013-7525.
Khatkar, M. S., G. Moser, B. J. Hayes, and H. W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13:538. http://dx.doi.org/10.1186/1471-2164-13-538.
Larmer, S. G., M. Sargolzaei, and F. S. Schenkel. 2014. Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. J. Dairy Sci. 97:3128–3141. http://dx.doi.org/10.3168/jds.2013-6826.
Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. 2010. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34:816–834. http://dx.doi.org/10.1002/gepi.20533.
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. http://dx.doi.org/10.3168/jds.2012-6316.
Marchini, J., and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11:499–511. http://dx.doi.org/10.1038/nrg2796.
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. http://dx.doi.org/10.3168/jds.2011-4490.
Neves, H. H., R. Carvalheiro, A. M. O. Brien, Y. T. Utsunomiya, A. S. do Carmo, F. S. Schenkel, J. Sölkner, J. C. McEwan, C. P. Van Tassell, J. B. Cole, M. V. da Silva, S. A. Queiroz, T. S. Sonstegard, and J. F. Garcia. 2014. Accuracy of genomic predictions in *Bos indicus* (Nellore) cattle. Genet. Sel. Evol. 46:17. http://dx.doi.org/10.1186/1297-9686-46-17.
Nicolazzi, E. L., S. Biffani, and G. Jansen. 2013. Short communication: Imputing genotypes using PedImpute fast algorithm combining pedigree and population information. J. Dairy Sci. 96:2649–2653. http://dx.doi.org/10.3168/jds.2012-6062.
Pausch, H., B. Aigner, R. Emmerling, C. Edel, K.-U. Götz, and R. Fries. 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet. Sel. Evol. 45:3. http://dx.doi.org/10.1186/1297-9686-45-3.
Pérez O'Brien, A. M., G. Mészáros, Y. T. Utsunomiya, T. S. Sonstegard, J. F. Garcia, C. P. Van Tassell, R. Carvalheiro, M. V. B. da Silva, and J. Sölkner. 2014. Linkage disequilibrium levels in *Bos indicus* and *Bos taurus* cattle using medium and high density SNP chip data and different minor allele frequency distributions. Livest. Sci. 166:121–132. http://dx.doi.org/10.1016/j.livsci.2014.05.007.
Porto-Neto, L. R., J. W. Kijas, and A. Reverter. 2014. The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genet. Sel. Evol. 46:22.
Pryce, J. E., J. Johnston, B. J. Hayes, G. Sahana, K. A. Weigel, S. McParland, D. Spurlock, N. Krattenmacher, R. J. Spelman, E. Wall, and M. P. L. Calus. 2014. Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations. J. Dairy Sci. 97:1799–1811. http://dx.doi.org/10.3168/jds.2013-7368.
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and

P. C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575. http://dx.doi.org/10.1086/519795.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. The R Foundation for Statistical Computing, Vienna, Austria.

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. http://dx.doi.org/10.1186/1471-2164-15-478.

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer. 2008. Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Sci. 91:2106–2117. http://dx.doi.org/10.3168/jds.2007-0553.

Schrooten, C., R. Dassonneville, V. Ducrocq, R. F. Brøndum, M. S. Lund, J. Chen, Z. Liu, O. González-Recio, J. Pena, and T. Druet. 2014. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip. Genet. Sel. Evol. 46:10. http://dx.doi.org/10.1186/1297-9686-46-10.

Utsunomiya, Y. T., L. Bomba, G. Lucente, L. Colli, R. Negrini, J. A. Lenstra, G. Erhardt, J. F. Garcia, and P. Ajmone-Marsan. 2014. Revisiting AFLP fingerprinting for an unbiased assessment of genetic structure and differentiation of taurine and zebu cattle. BMC Genet. 15:47. http://dx.doi.org/10.1186/1471-2156-15-47.

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

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

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. C. H. M. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, and G. A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 96:668–678. http://dx.doi.org/10.3168/jds.2012-5702.

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

Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, and D. B. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93:5942–5949. http://dx.doi.org/10.3168/jds.2010-3335.

Weigel, K. A., G. de Los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010a. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93:5423–5435. http://dx.doi.org/10.3168/jds.2010-3149.

Weigel, K. A., C. P. Van Tassell, J. R. O'Connell, P. M. VanRaden, and G. R. Wiggans. 2010b. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93:2229–2238. http://dx.doi.org/10.3168/jds.2009-2849.

Wiggans, G. R., T. A. Cooper, C. P. Van Tassell, T. S. Sonstegard, and E. B. Simpson. 2013. Technical note: Characteristics and use of the Illumina BovineLD and GeneSeek Genomic Profiler low-density bead chips for genomic evaluation. J. Dairy Sci. 96:1258–1263. http://dx.doi.org/10.3168/jds.2012-6192.

Wiggans, G. R., T. A. Cooper, P. M. Vanraden, K. M. Olson, and M. E. Tooker. 2012. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. J. Dairy Sci. 95:1552–1558. http://dx.doi.org/10.3168/jds.2011-4985.

Yu, X., J. A. Woolliams, and T. H. Meuwissen. 2014. Prioritizing animals for dense genotyping in order to impute missing genotypes of sparsely genotyped animals. Genet. Sel. Evol. 46:46 http://dx.doi.org/10.1186/1297-9686-46-46.

Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93:5487–5494. http://dx.doi.org/10.3168/jds.2010-3501.