Evaluation of developed low-density genotype panels for imputation to higher density in independent dairy and beef cattle populations

M. M Judge,* J. F. Kearney,†‡ M. C. McClure,† R. D. Sleator,‡ and D. P. Berry*2

*Teagasc, Animal & Grassland Research and Innovation Center, Moorepark, Fermoy, Co. Cork, Ireland; †Irish Cattle Breeding Federation, Highfield House, Bandon, Co. Cork, Ireland; and ‡Department of Biological Sciences, Cork Institute of Technology, Bishopstown, Co. Cork, Ireland

ABSTRACT: The objective of this study was to develop, using alternative algorithms, low-density SNP genotyping panels (384 to 12,000 SNP), which can be accurately imputed to higher-density panels across independent cattle populations. Single nucleotide polymorphisms were selected based on genomic characteristics (i.e., linkage disequilibrium [LD], minor allele frequency [MAF], and genomic distance) in a population of 1,267 Holstein–Friesian animals genotyped on the Illumina Bovine50 Beadchip (54,001 SNP). Single nucleotide polymorphism selection methods included 1) random; 2) equidistant location; 3) combination of SNP MAF and LD structure while maintaining relatively equal genomic distance between adjacent SNP; 4) a combination of high MAF, genomic distance between selected and candidate SNP, and correlation between genotypes of selected and candidate SNP; and 5) a machine learning algorithm. The panels were validated separately in 1) a population of 750 Holstein–Friesian animals with masked genotypes to reflect the lower-density SNP densities under investigation (1,249 animals with complete genotypes included in the reference population) and 2) a population of 359 Limousin and Charolais cattle with high (777,962 SNP)-density genotypes (1,918 animals with complete genotypes included in the reference population). Irrespective of SNP selection method, imputation accuracy in both populations improved at a diminishing rate as the number of SNP included in the lower-density genotype panel increased. Additionally, the variability in mean imputation accuracy per individual decreased as the panel density increased. The SNP selection method had a major impact on the mean allele concordance rate, although its impact diminished as the panel density increased. Imputation accuracy for SNP selected using a combination of high SNP MAF, LD structure, and relatively equal genomic distance between SNP outperformed all other selection methods in densities < 12,000 SNP. Using this method of SNP selection, the correlation between the imputed and actual genotypes for the 3,000 SNP panel was 0.90 and 0.96 when applied to the beef and dairy populations, respectively; the respective correlations for the 6,000 SNP panel were 0.95 and 0.98. It is necessary to include between 3,000 and 6,000 SNP in a low-density panel to achieve adequate imputation accuracy to either medium density (approximately 50,000 SNP in the dairy population) or high density (approximately 700,000 SNP in the beef population) across diverse and independent populations.

Key words: beef, dairy, genotype panels, imputation, single nucleotide polymorphism


INTRODUCTION

Dense genomic information is now being included in national dairy (Hayes et al., 2009; Spelman et al., 2013) and beef (Garrick, 2011) genetic evaluations to increase the accuracy of selection. The cost of procuring a genotype, even when using commer-
cially available low-density genotype panels, is imped-
ing widespread on-farm uptake of this technology. One option to reduce the cost of acquiring a genomically en-
hanced EBV for an animal is to use an even lower-den-
sity genotype panel, which can be imputed up to higher
density (Berry et al., 2013). This imputation should be
achieved with minimal loss in accuracy of the derived
 genomically enhanced EBV. Even slight reductions in
the cost of a genotype panel are likely to increase na-
tional uptake, which would, in turn, facilitate a greater
sized purchase order, thereby reducing the cost per unit
genotype. Such increased uptake can be best achieved
by designing a genotyping platform that is applicable
across breeds and populations, including those independ-
ent to the population used in generating the optimal
genotype panel (i.e., dairy versus beef populations).

The focus of the current study was to quantify the
accuracy of imputation achievable from lower-density
panels, derived using different algorithms to select the
most informative SNP. Because genotype panels may
sometimes be developed using information from just 1
population and potentially used in completely different
populations, the panels in the present study were de-
veloped based on genomic characteristics (i.e., linkage
disequilibrium [LD] patterns, SNP minor allele fre-
quency [MAF], and genomic distance between SNP)
from a sample of Holstein–Friesian animals and were
externally validated in both dairy and beef populations.
Results from this study will be useful in evaluating the
imputation accuracy achievable with genotype panels
of different densities and especially the transferability
of such panels across multiple populations that were
not actually used in the original design of the panel.
Such results could be used in subsequent cost–benefit
analyses of genomic prediction accuracy for different
cost panels with different imputation (and thus genomic-
ic prediction; Berry and Kearney, 2011) accuracies.

MATERIALS AND METHODS

Genotype Data

A total of 54,001 SNP on the Illumina Bovine50
Beachip (Matukumalli et al., 2009) were available
from 6,369 Holstein–Friesian dairy animals. Single
nucleotide polymorphism positions from the UMD3.1
(Zimin et al., 2009) and Btau4.0 (The Bovine
Genome Sequencing and Analysis Consortium, 2009)
genome builds were compared. To improve the con-
fidence in the reported genomic location of candi-
date SNP, SNP not recorded on the same autosome
in both builds were discarded as were SNP with an
unknown genomic position in either of the 2 builds.
Single nucleotide polymorphisms were then ranked,
within chromosome, by position from the start of the
chromosome, and SNP differing in position by more
than 50 rank points were discarded.

Mendelian inconsistencies were used to validate
animal identification through parentage; autosomal
SNP with >2% Mendelian inconsistancies between par-
ent–progeny pairs were discarded. Finally, SNP were
filtered on call rate (> 95%), MAF (> 2%), GenTrain
score (> 0.55; GenTrain scores were generated using
the GenCall software [Illumina Inc., San Diego, CA]),
and deviation from Hardy–Weinberg equilibrium ($P <
0.01 \times 10^{-9}$). Following edits, 40,483 autosomal SNP
remained. The Btau4.0 build was used for the remain-
der of this study.

Illumina HD genotypes (777,962 SNP; Illumina,
Inc.) were also available on 2,277 Limousin ($n = 1,380$)
and Charolais ($n = 987$) animals. Single nucleotide
polymorphisms with unknown positions or chromo-
somes were discarded. Duplicate SNP were also dis-
carded. After edits, 735,239 SNP remained.

Animal Populations

Three distinct animal populations were generated
from the 6,369 Holstein–Friesian animals: 1) a popula-
tion with complete genotype information to estimate
LD patterns and SNP MAF for SNP selection, 2) a (reference) population with complete genotype infor-

mation to estimate haplotypes for SNP imputation, and
3) a (validation) population with masked genotypes
reflective of the lower-density panels. The validation
population consisted of 750 animals and comprised
30 of the youngest Holstein–Friesian animals (with both
sire and dam genotyped), their paternal half sibs, and
any animals that shared a maternal grandsire with them.

A total of 2,868 animals whose sire, dam, or ma-
ternal grand-sire appeared either in the validation
population itself or in the recent 2 generations of the
validation population were not further considered in
the study. The remaining 2,751 animals were used for
the estimation of both LD patterns and SNP MAF ($n = 1,267$) and as the reference population for haplotype
generation ($n = 1,484$).

The youngest 212 Limousin and 147 Charolais
animals were selected as the beef external validation
population; 1,918 Limousin ($n = 1,168$) and Charolais
($n = 750$) animals were included in the imputation ref-

erence population.

Development of the SNP Panels

Low-density genotype panels were generated to
reflect 6 different SNP densities: 384, 1,000, 2,000,
3,000, 6,000, and 12,000 SNP. Six different algorithms
Low-density genotype panels

were used to select the informative SNP. The number of SNP selected per chromosome remained constant for each SNP selection algorithm evaluated. The number of SNP per chromosome required for each density was directly proportional to the genomic length of each chromosome and is summarized for each density in Table 1. The 6 alternative algorithms used to generate each of the SNP densities were as follows:

1) Random SNP Selection Method – Single nucleotide polymorphisms, within each chromosome, were selected at random until the predefined threshold number of SNP per chromosome for the panel density being developed was reached.

2) Uniform SNP Selection Method – Single nucleotide polymorphisms were selected uniformly across chromosomes. The increment between SNP was the same for all chromosomes for a given panel density to achieve, as close as possible, the predefined number of SNP per chromosome for that density (Table 1). Where the number of SNP was greater than the desired number, SNP in the center of the chromosome (largest chromosomes first) were discarded with a maximum discarding of 1 SNP per chromosome.

3) Block SNP Selection Method – This SNP selection algorithm divided each chromosome into blocks of SNP, with 1 representative SNP chosen per block. The number of blocks per chromosome was equal to the predefined number of SNP for that chromosome minus 2; this was undertaken so an extra SNP could be selected on the peripheries (i.e., the first 0.5 Mb) of the chromosome. Single nucleotide polymorphisms were ranked on an index of the MAF of that SNP plus the average LD between that SNP and all other candidate SNP in that block; both MAF and average LD were standardized to have an equal variance before summing, so an equal weight was placed on both attributes. The highest-ranking SNP (i.e., high MAF and strong average LD) was chosen within each block. For the blocks on the immediate periphery of each chromosome, a second informative SNP was also selected. The partial correlation of each candidate SNP in the block with all other remaining SNP in that block (standardized to have equal variances) was calculated as

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

in which \(r(SNP_i', SNP_j'|SNP_{sel})\) is the correlation of candidate SNP (\(SNP_i'\) and \(SNP_j'\)) after adjusting for the relationship of these SNP with the selected SNP (\(SNP_{sel}\)).

The highest-ranking SNP on an index of MAF plus the average partial correlations between that SNP and all other remaining SNP in that block (standardized to have equal variances) was chosen as the second most informative SNP.

4) Wellmann SNP Selection Method (Wellmann et al., 2013) – Single nucleotide polymorphisms were selected based on their genomic location, high SNP MAF, and weak correlation between the genotypes of already selected and remaining candidate SNP as outlined in detail by Wellmann et al. (2013).

5) Modified Wellmann SNP Selection Method – This method is an expansion of the previously described Wellmann SNP selection method (Wellmann et al., 2013). In addition to selecting a candidate SNP that had a weak correlation with the already selected SNP, SNP with a strong correlation with the remaining candidate SNP were also prioritized. This was achieved by generating a pairwise distance score,
within chromosome, between SNP. The correlation between each candidate SNP and all remaining SNP in that chromosome was also calculated. An index value was then generated for each candidate SNP based on the sum of SNP MAF, distance score, and average LD with remaining candidate SNP; the variance of each of the 3 attributes were standardized to be equal. The highest index value SNP was then selected and the process was iterated until the quota of SNP per chromosome had been reached for the SNP density panel under investigation.

6) Feature Selection using Feature Similarity (FSFS) – This machine learning method of SNP selection involves grouping features (i.e., SNP) into clusters so that SNP within each cluster are similar (i.e., strong LD); a single SNP within each cluster is then chosen to represent the cluster (Phuong et al., 2006). The algorithm is outlined in detail in Appendix 1.

Imputation

Imputation to the higher-density panels was undertaken (separately for both populations) using both FImpute version 2.2 (Sargolzaei et al., 2014) exploiting family- and population-based information and Beagle version 3.3.2 (Browning and Browning, 2007, 2009). For FImpute, all chromosomes were simultaneously imputed, whereas for Beagle, chromosomes were separately imputed. Ten generations of pedigree were used in the imputation process for FImpute whereas 1 generation of pedigree was used in the imputation process for Beagle. Preliminary analyses revealed minimal impact on imputation accuracy when only 1 generation of pedigree was used in FImpute. Imputation for all scenarios was undertaken within breed.

Not all SNP on the Illumina Bovine50 Beadchip (used in the dairy population) are on the Illumina High Density genotype platform (used in the beef population; Illumina, Inc.). If such SNP were selected for the lower-density platform, then the nearest SNP on the Illumina High Density platform was selected for the imputation in the beef population.

The association between chromosome length and imputation accuracy was determined using a fixed effects model where the dependent variable was the mean allele concordance rate per chromosome. How the association differed by SNP density and SNP selection algorithm was also evaluated by including these main effects and their interaction with chromosome length as fixed effects in the model. Significance was based on the $F$ statistic.

RESULTS

Dairy Validation Population

Imputation Algorithms. The difference in imputation accuracy between FImpute and Beagle was minimal and diminished as the panel density increased. Across all SNP selection algorithms, Beagle outperformed FImpute in imputation accuracy for panel densities of ≤2,000 SNP (Fig. 1; Table 2). For example, when SNP were selected using the block method, Beagle achieved, on average, 0.013, 0.005, and 0.0004 greater mean allele concordance rates than FImpute for the 384, 1,000, and 2,000 SNP density panels, respectively. However, for higher-density panels, imputation accuracy for FImpute outperformed Beagle. The panel density at which the crossover in accuracy of both imputation algorithms occurred was dependent on the SNP selection algorithm. For both the block and Wellmann SNP selection methods, the crossover occurred at 3,000 SNP. When SNP were selected using the uniform, random, and FSFS methods, the crossover occurred at the 6,000 SNP density. When SNP were selected using the block selection method, the difference in the mean allele concordance rate between FImpute and Beagle varied from 0.001 (12,000 SNP) to 0.002 (6,000 SNP). The same trend in imputation accuracy differences between the 2 imputation algorithms also existed when the correlation was used as the measure of imputation accuracy.

The frequency the same genotype was imputed correctly from the 3,000 SNP density panel (i.e., both alleles correct) by both imputation algorithms was 92.37% (Table 3). Both FImpute and Beagle incorrectly imputed the same genotype (i.e., no allele correct) 0.01% of the time, which represented 16% of the incorrect genotypes for FImpute and 23% of the incorrect genotypes for Beagle.

The variability in the mean allele concordance rate per animal was less with FImpute than with Beagle, regardless of the panel density or SNP selection algorithm (Fig. 1). For example, when SNP were selected with the block method, imputation with FImpute resulted in a 30% less variable allele concordance rate per animal on the 3,000 SNP panel compared with Beagle.

Beagle was capable of imputing minor alleles with marginally greater accuracy than FImpute (a difference in the mean allele concordance rate of 0.0007 when MAF was ≤0.05 and a difference of 0.001 when MAF was between 0.051 and 0.10; Fig. 2). In addition to this, less variability in the allele concordance rate across MAF bins was evident for Beagle. Because of the negligible difference in the accuracy of imputation (expressed as either the concordance rate or corre-
tion) between FImpute and Beagle (Table 2), all results from hereon in will refer only to FImpute.

Panel Density. The mean animal allele concordance rate and the correlation between actual and imputed genotypes per animal improved at a diminishing rate as the panel density increased, regardless of the SNP selection algorithm used (Fig. 3). With the block method of SNP selection, for example, the mean animal allele concordance rate increased by 3% when the SNP density doubled from 1,000 to 2,000. Doubling the SNP density from 3,000 to 6,000 and from 6,000 to 12,000 increased the mean animal allele concordance rate by only 1 and 0.6%, respectively; the respective increases in the correlation between actual and imputed genotypes were 7, 2, and 1%. Moreover, regardless of the SNP selection method, the variability in the mean allele concordance rate per animal reduced as SNP density increased (Fig. 3). For example, with 384 SNP selected using the block method, the mean animal allele concordance rate was 0.849 (minimum 0.730 and maximum 0.980), whereas with the 12,000 SNP panel, the mean animal allele concordance was 0.994 (minimum 0.943 and maximum 0.999). Similarly, using the uniform method of SNP selection, the mean

Table 2. The allele concordance rate (ACR) and correlation (r) between actual and imputed genotypes for different low-density SNP panels, different SNP selection algorithms, and different imputation algorithms

<table>
<thead>
<tr>
<th>SNP selection method</th>
<th>Impute method</th>
<th>384</th>
<th>1,000</th>
<th>2,000</th>
<th>3,000</th>
<th>6,000</th>
<th>12,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>FImpute¹</td>
<td>0.849</td>
<td>0.928</td>
<td>0.963</td>
<td>0.976</td>
<td>0.988</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>Beagle²</td>
<td>0.862</td>
<td>0.932</td>
<td>0.964</td>
<td>0.975</td>
<td>0.986</td>
<td>0.993</td>
</tr>
<tr>
<td>Wellmann</td>
<td>FImpute</td>
<td>0.833</td>
<td>0.922</td>
<td>0.961</td>
<td>0.974</td>
<td>0.987</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>Beagle</td>
<td>0.845</td>
<td>0.927</td>
<td>0.962</td>
<td>0.974</td>
<td>0.985</td>
<td>0.992</td>
</tr>
<tr>
<td>Uniform</td>
<td>FImpute</td>
<td>0.830</td>
<td>0.907</td>
<td>0.950</td>
<td>0.966</td>
<td>0.984</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>Beagle</td>
<td>0.840</td>
<td>0.914</td>
<td>0.953</td>
<td>0.967</td>
<td>0.982</td>
<td>0.991</td>
</tr>
<tr>
<td>Random</td>
<td>FImpute</td>
<td>0.834</td>
<td>0.905</td>
<td>0.946</td>
<td>0.963</td>
<td>0.981</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td>Beagle</td>
<td>0.847</td>
<td>0.912</td>
<td>0.949</td>
<td>0.964</td>
<td>0.980</td>
<td>0.989</td>
</tr>
<tr>
<td>FSFS³</td>
<td>FImpute</td>
<td>0.823</td>
<td>0.888</td>
<td>0.940</td>
<td>0.957</td>
<td>0.979</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td>Beagle</td>
<td>0.835</td>
<td>0.896</td>
<td>0.946</td>
<td>0.961</td>
<td>0.979</td>
<td>0.989</td>
</tr>
</tbody>
</table>

¹Sargolzaei et al. (2014).
²Browning and Browning (2007, 2009).
³FSFS = Feature Selection using Feature Similarity.
animal allele concordance rate for the 384 SNP panel was 0.830 (minimum 0.721 and maximum 0.974), whereas with the 12,000 SNP panel, the mean animal allele concordance rate was 0.992 (minimum 0.933 and maximum 0.999).

**Single Nucleotide Polymorphism Selection Algorithm.** The SNP selection algorithm had a major impact on the accuracy of imputation to higher density, regardless of the SNP density of the lower-density platform (Table 2; Fig. 3). Imputation accuracy from SNP selected using the block method outperformed all other SNP selection algorithms in panel densities 384, 1,000, 2,000, 3,000, and 6,000 (0.016, 0.0002, 0.001, 0.002, and 0.001 greater allele concordance rate and 0.032, 0.009, 0.003, 0.005, and 0.001 stronger correlation with the true genotypes when compared with the next best SNP selection algorithm). The FSFS SNP selection algorithm consistently resulted in the poorest imputation accuracy (both the mean animal allele concordance rate and the correlation between actual and imputed genotypes) across all panel densities, with the exception of the 384 SNP density panel (Table 2; Fig. 3). When the panel density reached 12,000 SNP, the effect of the SNP selection algorithm on imputation accuracy was negligible; the mean allele concordance rate per animal for each SNP selection algorithm varied from 0.985 to 0.990.

The SNP selection algorithm also affected the variability in the mean allele concordance rate per animal (Fig. 3). Single nucleotide polymorphism selection algorithms that achieved the greatest mean imputation accuracy exhibited the least variability in the mean allele concordance rate per individual. When SNP were selected using the block method, the mean allele concordance rate with the 3,000 SNP panel was 0.976 with a SD of 0.1, whereas for the same density panel, SNP selected with the uniform method had a mean animal allele concordance rate of 0.966 but with a SD of 0.12.

With the exception of the 384 SNP panel, the accuracy of imputation was marginally lower for the modified Wellmann method than that achieved using the original Wellmann method. For example, when SNP were selected with the modified Wellmann method on the 3,000 SNP density panel, the mean allele concordance rate per animal decreased from 0.974 to 0.972 and the correlation between actual and imputed genotypes decreased from 0.959 to 0.955.

**Beef Validation Population**

Only SNP selected using the block method or the Wellmann method were considered for validation in the beef population. The block method of SNP selection marginally outperformed the Wellmann method, regardless of SNP panel density (an improvement of

| Table 3. Contingency table of number of alleles correctly imputed on a 3,000 SNP density panel selected using the block method for FImpute\(^1\) and Beagle\(^2\) |
|-----------------------|------------------|------------------|------------------|
|                       | FImpute          | Beagle           |
|-----------------------|------------------|------------------|------------------|
| 0 alleles correct     | 0.01             | 0.03             | 0.04             |
| 1 allele correct      | 0.02             | 2.00             | 2.62             |
| 2 alleles correct     | 0.02             | 2.89             | 92.37            |

\(^1\)Sargolzaei et al. (2014).

\(^2\)Browning and Browning (2007, 2009).
Regardless of SNP selection algorithm, both the mean allele concordance rate per animal and the mean correlation between actual and imputed genotypes increased at a diminishing rate as the SNP density increased (Fig. 4). Regardless of the SNP selection method, the variability in both the allele concordance rate per animal and the correlation between actual and imputed genotypes was reduced as the SNP density increased (Fig. 4). For example, with the 384 SNP panel, using SNP selected by the block method, the mean animal allele concordance rate was 0.789 (minimum 0.734 and maximum 0.869), whereas with the 12,000 SNP panel, the mean value was 0.985 (minimum 0.894 and maximum 0.996). Similarly, using the Wellmann method of SNP selection, the mean animal allele concordance rate on the 384 SNP panel was 0.786 (minimum 0.735 and maximum 0.861), whereas on the 12,000 SNP panel, the mean value was 0.986 (minimum 0.894 and maximum 0.996).

**DISCUSSION**

The motivation for this study was to quantify the accuracy of imputation to medium- (i.e., approximately 50,000 SNP; dairy population) and high-density (i.e., approximately 777,000 SNP; beef population) genotype platforms from custom-derived, lower-density panels across populations. Reducing the SNP density should result in a reduced cost per unit genotype. This must be achievable without a large compromise in accuracy of generating the higher-density genotypes for use in genomic predictions or genome-wide association studies. The most pertinent research questions of the present study were first, what is the sacrifice in accuracy of imputation from low-density SNP panels of differing densities, and second, how best to select SNP for inclusion on lower-density panels. Both questions have been addressed in the current study. Also of particular interest in the present study was the transferability of panels across independent populations; commercial genotype panels are often developed using the genomic characteristics of a single or a few populations and then applied in a range of other (sometimes unrelated) populations. Therefore, the accuracy of imputation achieved in the present study should be considered the lower threshold and could possibly be improved if information from the population in which the panel was to be applied was also included in the development of the panel.

Several studies have evaluated the accuracy of imputation from alternative-density genotype panels ranging from commercially available panels (Berry and Kearney, 2011; Berry et al., 2014) to custom-derived panels of SNP (Habier et al., 2009; Weigel et al.,...
The choice of informative SNP to include on such panels is likely to be a key factor in ensuring adequate accuracy of imputation from low- to high-density genotype panels. Previously evaluated approaches in cattle to select SNP for lower-density panels include random selection (Szyda et al., 2013), uniform selection across the genome (Habier et al., 2009), a combination of equidistant physical location and high MAF (Weigel et al., 2010; Boichard et al., 2012), dividing each chromosome into equal sized segments and selecting the SNP closest to the physical midpoint of each segment or selecting the SNP with the greatest MAF within each segment (Mulder et al., 2012), or selecting SNP based on pairwise LD patterns (Szyda et al., 2013). Further SNP selection algorithms have been proposed in other species. Wellmann et al. (2013), for example, described a method of SNP selection in swine based on a combination of MAF, genomic distance between selected and candidate SNP, and the absolute correlation between alleles of selected and candidate SNP. Approaches such as machine learning algorithms have also been proposed for the selection of SNP (Phuong et al., 2006).

Several studies have documented the accuracy of imputation from low- to high-density panels using real-life cattle (Chud et al., 2014; Boichard et al., 2012; Weigel et al., 2010; Druet and Georges, 2010; Zhang and Druet, 2010), pig (Wellmann et al., 2013), equine (Corbin et al., 2014), and sheep (Hayes et al., 2011) populations. It is desirable that any newly developed low-density panels be applicable across breeds and populations, especially if crossbreeding exists, as is the case in Ireland (Berry et al., 2006). The mean imputation accuracy for the different density panels in the present study was generally consistent with imputation accuracies documented for comparable-density panels in other populations of cattle (Table 4). The number of SNP in the higher-density panel in the dairy population in the present study (approximately 50,000 SNP) was, however, lower than the higher-density panel in the beef population (approximately 700,000 SNP).

**Imputation Algorithm**

Beagle and FImpute are imputation software packages commonly used for cattle (Berry and Kearney, 2011; Berry et al., 2014; Carvalheiro et al., 2014). Beagle uses population LD to impute missing genotypes whereas FImpute exploits both family information and population LD. The marginal improvement in imputation accuracy when using FImpute over Beagle (with ≥3,000 SNP) and the reduction of the variability in imputation accuracy corroborates previous studies that compared both software suites. Chud et al. (2014) reported an overall increase in imputation accuracy of 5 percentage units for FImpute over Beagle when imputing from 9,000 SNP to a high-density genotype panel consisting of 777,962 SNP in cattle. Similarly, Carvalheiro et al. (2014) documented an increase in average accuracy of 3.4% for FImpute over Beagle using a 7,000 SNP density panel to impute to a high-density panel of approximately 777,000 SNP. Furthermore, Chud et al. (2014) and Carvalheiro et al. (2014) both reported a reduction in the variability of the mean im-

![Figure 4](image-url)
## Table 4. Summary of imputation accuracy from low-density SNP panels to higher-density panels in cattle populations

<table>
<thead>
<tr>
<th>Number of animals</th>
<th>SNP select</th>
<th>Density</th>
<th>ACR</th>
<th>r</th>
<th>ACR</th>
<th>r</th>
<th>ACR</th>
<th>r</th>
<th>ACR</th>
<th>r</th>
<th>ACR</th>
<th>r</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val</td>
<td>Cal</td>
<td>Alg</td>
<td>384</td>
<td>&gt;384 ≤ 1,000</td>
<td>&gt;1,000 ≤ 2,000</td>
<td>&gt;2,000 ≤ 3,000</td>
<td>&gt;3,000 ≤ 6,000</td>
<td>&gt;6,000 ≤ 12,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>764</td>
<td>4,732</td>
<td>B</td>
<td>Com</td>
<td>0.6–0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,149</td>
<td>793</td>
<td>FI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>283</td>
<td>FI</td>
<td>Add</td>
<td>0.46–0.53</td>
<td>0.64–0.78</td>
<td>0.86–0.88</td>
<td>0.88–0.94</td>
<td>0.93–0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>971</td>
<td>2,931</td>
<td>B</td>
<td>Com</td>
<td>0.96–0.97</td>
<td>0.83–0.90</td>
<td>0.94–0.97</td>
<td>0.85–0.92</td>
<td>0.91–0.98</td>
<td>0.66–0.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8,145</td>
<td></td>
<td>FI</td>
<td>Com</td>
<td>0.16–0.81</td>
<td>0.57–0.96</td>
<td>0.61–0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4,734</td>
<td>500–2,000</td>
<td>DP</td>
<td>Mid SNP in group</td>
<td>0.14–0.78</td>
<td>0.56–0.96</td>
<td>0.61–0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,425</td>
<td>4,822</td>
<td>B</td>
<td>High MAF in group</td>
<td>0.49–0.71</td>
<td>0.56–0.75</td>
<td>0.61–0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>698</td>
<td>2,424</td>
<td>B</td>
<td>Com</td>
<td>0.96–0.98</td>
<td>0.94–0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 Val = validation population size for imputation; Cal = calibration (reference) population size for imputation.
2 Alg = imputation algorithm used. Imputation software used. B = Beagle (Browning and Browning, 2007, 2009); FI = FImpute (Sargolzaei et al., 2014); O = Alphaimpute (Hickey et al., 2012) or findhap (VanRaden et al., 2011); DP = DAGPHASE (Druet and Georges, 2010); CH = CHROMBID (Druet and Farnir, 2011).
3 SNP select = SNP selection method; Com = SNP were selected and added to a base panel from commercial panels; Add = a commercial panel base was used and SNP from other panels were included to form higher-density panels; High MAF in group = SNP were selected based on high minor allele frequency (MAF) within a bin; Mid SNP in group = SNP were selected when close to the physical midpoint of a bin; Equal + high MAF = SNP were selected on both high MAF and at an equal distance across the chromosome.
4 ACR = allele concordance rate.
5 The correlation between actual and imputed genotypes.
utation accuracy per individual when imputation was undertaken using FImpute compared with when imputation was undertaken using Beagle. A similar imputation accuracy but shorter computing time for FImpute over Beagle (present study; Sargolzaei et al., 2014; Chud et al., 2014) means that FImpute may be considered a more suitable imputation software package for use in cattle.

**Single Nucleotide Polymorphism Selection and Panel Density**

The trend of improved imputation accuracy with an increase in the panel density, across populations, is not unexpected and is similar to previous reports in cattle (Mulder et al., 2012; Carvalheiro et al., 2014), sheep (Hayes et al., 2011), equines (Corbin et al., 2014), and pigs (Wellmann et al., 2013). In both the dairy and beef populations, the conclusion reached in the present study, that doubling the panel density from 1,000 to 2,000 SNP resulted in a greater improvement in the mean allele concordance than doubling the panel density from 3,000 to 6,000 and especially from 6,000 to 12,000 SNP, was due to the fact that the allele concordance rate was already high at densities of ≥3,000 SNP (Fig. 3). This trend of diminishing benefit in imputation accuracy with increasing the panel density is consistent with results outlined by Carvalheiro et al. (2014), where an increase in the panel density from 7,000 to 15,000 SNP generated a greater improvement in mean imputation accuracy (8% improvement) than an increase in the panel density from 15,000 to 75,000 SNP (1% improvement). Another reason for this trend may be that the higher-density panel has fewer masked SNP. As known SNP were included in the calculation of imputation accuracy in the present study, a slightly inflated (yet representative of reality) accuracy statistic of actual imputation accuracy may result, when compared with studies that did not include the unmasked genotypes in their accuracy calculations (Carvalheiro et al., 2014). The reduction in the mean allele concordance rate when focusing only on the masked SNP is in Appendix 2. Therefore, basing the allele concordance rate on just the unmasked genotypes does not alter the conclusion of a diminishing benefit in imputation accuracy as density increased.

Although mean imputation accuracy per individual is an important measure of accuracy, variability in mean imputation accuracy per individual is also extremely important. Individual breeders and producers are concerned with not only the average accuracy of imputation for the entire population but, more importantly, the imputation accuracy of their individual animals. The phenomenon of a reduction in variability in the mean allele concordance rate as the panel density increased has been previously documented in cattle (Judge et al., 2014). Carvalheiro et al. (2014) stated that there was a reduction of 0.022 in the SD of the mean imputation accuracy when increasing SNP density from 7,000 to 11,000 SNP. There was a further reduction of 0.001 in the SD of the mean imputation accuracy when the SNP density increased from 11,000 to 15,000 SNP.

Corbin et al. (2014) quantified the impact of imputation accuracy using several developed low-density panels (384, 768, 1,000, 2,000, 3,000, and 6,000 SNP) in Thoroughbred horses from different geographical locations. Single nucleotide polymorphisms included in their proposed low-density panels were selected using 1 of 3 methods: 1) approximately equidistant across the genome, 2) approximately equidistant across the genome and optimized for MAF, and 3) optimized for both MAF and LD units, the latter as a proxy for genomic distance and recombination rates. The low-density panels, developed using 853 U.K. Thoroughbred horses, were used to impute 348 U.S. Thoroughbred horses to approximately 50,000 SNP. Corbin et al. (2014) recommended that at least 2,000 SNP were necessary for accurate imputation (a mean proportion of correctly imputed genotypes of 0.91). Similarly, in the present study, a genotype concordance rate (i.e., the proportion of correctly imputed genotypes) of 0.92 was achieved with a low density of 2,000 SNP selected using the block method. This was despite the LD between SNP 50 kb apart being approximately 0.4 in Thoroughbred horses (Corbin et al., 2010), which is double that observed in cattle (Porto-Neto et al., 2014).

Nonetheless, at least 3,000 SNP are required to achieve high imputation accuracy (mean allele concordance rate of ≥0.976, SD of ≤0.013, and a minimum individual mean allele concordance value of 0.857) in the dairy population; the requirement of at least 3,000 SNP for low-density panel is in agreement with other studies undertaken in cattle (Mulder et al., 2012; Zhang and Druet, 2010). In order for the low-density panels to be applicable to beef populations, however, which are completely independent of the dairy population, at least 6,000 SNP were required to achieve adequate imputation accuracy (mean allele concordance rate of ≥0.97, SD of ≤0.166, and a minimum individual mean allele concordance value of 0.842). This threshold could possibly be reduced if beef animals were also included in the selection of SNP for inclusion in the low-density panel. To further investigate this scenario, a 3,000 SNP density genotype panel was developed using genomic structure from only the beef population. Single nucleotide polymorphisms were selected using the block method of SNP selection, using 1,918 (beef) animals as the reference population; the 359 youngest Charolais
and Limousin animals were again used as the validation population. The mean individual animal genotype and the allele concordance rate for the validation population was 0.916 (SD 0.044) and 0.959 (SD 0.024), respectively. Therefore, the accuracy of imputation was, as expected, improved when animals representative of the population in which the low-density genotype platform is to be used were considered in the development of the panel itself, but also, the individual animal variability was reduced when the population in which the panel will be applied was also used in the generation of the panel.

The method of SNP selection influenced imputation accuracy in the present study, although the impact tended to diminish as SNP density increased. The FSFS SNP selection method resulted in the worst accuracy of imputation across panel densities and populations and, therefore, is not a practical method of choosing SNP for use in genomic selection, especially for low-density panels. Carvalheiro et al. (2014) reported that SNP selected based on a combination of MAF and LD yielded better imputation accuracy when compared to selection methods that used just one of these methods. In the present study, both the block and Wellmann methods of SNP selection placed an equal emphasis on both high MAF and LD when selecting SNP for use on the low-density panels. The reason the block method of SNP selection achieved superior imputation accuracy may be due to the positioning of SNP across the genome. Single nucleotide polymorphisms selected with the block method were forced to be more evenly distributed across the genome because only 1 SNP per segment could be selected; the Wellmann method, however, allowed adjacent SNP to be selected if the LD between them was low (and SNP in other regions of the chromosome had already been selected). The SD in distance between adjacent SNP was 510,362 and 868,746 bp for the block and Wellmann selection methods, respectively. Altering the K parameter of the Wellmann method did, however, also influence the variability in the position of SNP selected; the SD in distance between adjacent SNP was 384,342 bp for the uniform selection method, implying that genomic location of the selected SNP alone is not the only factor influencing the efficiency of the lower-density panels. Mean SNP MAF for the 3,000 SNP density panel for both the Wellmann (0.44) and block (0.46) methods was greater than that for the uniform (0.27), random (0.27), or FSFS (0.21) selection methods; this was attributable to the respective SNP selection algorithms with the former 2 SNP selection algorithms placing 50% emphasis on MAF, thus achieving near the maximum possible MAF of 0.5.

One reason for the lower allele concordance rate achieved in the beef population was that the SNP on the low-density panels were optimized based on the genomic characteristics in the dairy population, which may not be directly transmissible to other species. The mean SNP MAF in the Limousin and Charolais populations when SNP were selected using the block method were 0.23 and 0.23, respectively, and the respective SNP MAF for SNP selected using the Wellmann method were 0.23 and 0.23.

**Imputation Accuracy per SNP and per Chromosome**

The trend in the declining mean allele concordance rate with an increase in SNP MAF (Fig. 2) has been documented elsewhere (Berry and Kearney, 2011; Hayes et al., 2011). Corbin et al. (2014) reported that SNP with lower MAF tended to exhibit greater variability in imputation accuracy (i.e., the correlation and proportion of correctly imputed alleles). In the present study, however, the opposite was observed, and SNP that had low MAF also had the least variable allele concordance rates (Fig. 2).

The trend of increasing mean imputation accuracy per chromosome with increasing chromosome length for the lower-density panels in the present study (Fig. 5) corroborates other studies in cattle (Berry and Kearney, 2011) and horses (Corbin et al., 2014), although the association was dependent on both the SNP panel density ($P < 0.001$) and the SNP selection algorithm ($P < 0.001$). For example, using SNP selected by the block method on the 1,000 SNP density panel, every 100 Mb increase in chromosome length was associated ($P < 0.001$) with an increase in the mean allele concordance rate of 0.035 (SE 0.002; Fig. 6). As the panel density increased, the association between chromosome length and imputation accuracy weakened; with 12,000 SNP selected using the block method, the association between imputation accuracy of chromosome length was negative but close to 0 ($–0.009$; SE 0.002, $P < 0.001$).

Berry and Kearney (2011) reported that the mean allele concordance rate per chromosome varied from 0.964 to 0.979 when imputing from a commercially available 3,000 SNP panel to 50,000 SNP in Holstein–Friesian cattle. Sun et al. (2012) and Boichard et al. (2012) reported that imputation was more difficult at the peripheries of chromosomes due to a lack of flanking information present. Therefore, short chromosomes were at a disadvantage, because the poorly imputed peripheral chromosomal regions make up a larger proportion of the overall chromosome. Sun et al. (2012) investigated imputation in Angus cattle and considered...
Judge et al.

3 chromosomes in their study, chromosome 1 (representing long chromosomes), chromosome 16 (representing moderately sized chromosomes), and chromosome 28 (representing short chromosomes). Similar to the present study, both Sun et al. (2012) and Boichard et al. (2012) also selected extra SNP markers at the peripheries of each chromosome. Despite this, in both the present study and that of Sun et al. (2012), imputation accuracy persisted in being inferior in shorter chromosomes. Wellmann et al. (2013) also documented a relationship between chromosome length and imputation accuracy when using a genotyping panel containing 384 SNP in a pig population. However, unlike the current study, Wellmann et al. (2013) did not select extra SNP at the peripheries of chromosomes, which may have further disadvantaged the shorter chromosomes.

**Conclusion**

Accurate imputation is achievable with low-density genotype panels (i.e., 3,000 to 6,000 SNP) for both dairy and beef populations, once pedigree information is available. A selective SNP selection algorithm is, however, critical to achieve adequate imputation accuracy levels. The block SNP selection algorithm, which considers a combination of both SNP MAF and LD

---

**Figure 5.** The mean allele concordance rate per chromosome for the 3,000 SNP density panel, selected using the block method and imputed with FImpute (Sargolzaei et al., 2014).

**Figure 6.** The regression coefficient of the allele concordance rate on chromosome length (in 100-Mb units) for the dairy population when SNP were selected using the Wellmann (double black line), the block (dashed black line), the uniform (solid black line), the random (solid gray line), or the Feature Selection using Feature Similarity (dashed gray line) methods.
Appendix 1. The FSFS algorithm used for the machine learning method of SNP selection.

Input: \( S(F_1, F_2, ..., F_N) \)  
\( k \) \((k \leq N - 1)\)  
// originally SNP set \( S \)

Output: \( R \)  
// a tSNP subset \( R \)

1. \( R \leftarrow S \)
2. for each \( F_i \in R \) do
   \( d^i_k = D(F_i, F^k_i) \) where \( F^k_i \) is the \( k \)-th nearest neighbour of \( F_i \) in \( R \).
end for
3. find \( F_0 \) such that \( d^k_0 = \arg \min_{F_i \in R} d^k_i \)
   Let \( F^1_0, ..., F^k_0 \) be the \( k \) nearest SNPs of \( F_0 \)
   \( R \leftarrow R / \{ F^1_0, ..., F^k_0 \} \)
   if first iteration then set \( \Theta = d^k_0 \)
4. if \( k > |R| - 1 \) then \( k = |R| - 1 \)
5. if \( k = 1 \) goto 8.
6. while \( d^k_0 > \Theta \) do
   \( k = k - 1 \)
   if \( k = 1 \) goto 8
   recomputed \( d^k_0 \).
end while
7. goto 2
8. return \( R \)

Appendix 2. Reduction in mean allele concordance rate when unmasked SNPs were not included in the calculation of allele concordance rate across the six single nucleotide polymorphism densities and the four SNP selection methods

<table>
<thead>
<tr>
<th>Panel density</th>
<th>Random</th>
<th>Uniform</th>
<th>Wellman</th>
<th>Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>384</td>
<td>0.00159</td>
<td>0.00163</td>
<td>0.00160</td>
<td>0.00144</td>
</tr>
<tr>
<td>1,000</td>
<td>0.00241</td>
<td>0.00235</td>
<td>0.00198</td>
<td>0.00183</td>
</tr>
<tr>
<td>2,000</td>
<td>0.00282</td>
<td>0.00259</td>
<td>0.00201</td>
<td>0.00190</td>
</tr>
<tr>
<td>3,000</td>
<td>0.00299</td>
<td>0.00271</td>
<td>0.00205</td>
<td>0.00192</td>
</tr>
<tr>
<td>6,000</td>
<td>0.00333</td>
<td>0.00284</td>
<td>0.00220</td>
<td>0.00212</td>
</tr>
<tr>
<td>12,000</td>
<td>0.00378</td>
<td>0.00316</td>
<td>0.00252</td>
<td>0.00257</td>
</tr>
</tbody>
</table>

structure relatively equally positioned across the genome, provided the most accurate imputation accuracy.

LITERATURE CITED


