Comparing strategies for imputing informative sequence variants into a wider population of genotyped dairy cattle

K.M. Tiplady¹, R.G. Sherlock¹

¹ Livestock Improvement Corporation, Ruakura Rd, Newstead, Hamilton, New Zealand
ksanders@lic.co.nz (Corresponding Author)

Summary

Inclusion of well-chosen sequence variants in genomic prediction is a promising approach to improving accuracy. Sequence variants of interest in the implementation of this approach must be either genotyped directly or imputed using existing genotypes. Imputation is an attractive approach because it is less expensive than directly genotyping all the variants of interest. The purpose of this study was to quantify accuracy and computation times for a number of different imputation strategies in a mixed-breed set of genotyped dairy cattle. Of particular interest was whether accuracy for imputing a subset of sequence variants can be improved by using a large reference of ~11k genotyped individuals with imputed sequence instead of a small reference of ~500 individuals with true sequence. Whole-genome sequence for 563 individuals was used to quantify accuracy for imputing from ~8k density up to a subset of ~158k sequence variants which had previously been identified as informative for genomic prediction across traits of interest. Imputation was carried out using Beagle and FImpute, and accuracies for stepwise approaches via intermediate Illumina BovineSNP50k and BovineHD panels were also assessed. Forty-nine of the 563 individuals were used as a validation set. Strategies using Beagle and imputing to the larger reference of 11,457 individuals with imputed sequence vs imputing to the smaller reference of 514 individuals with true sequence were 0.962 and 0.946 respectively. Computation times were 8.7 and 0.03 CPU hours respectively. Imputation accuracies using FImpute and imputing to the larger reference with imputed sequence vs imputing to the smaller reference with true sequence were 0.953 and 0.952 respectively. Corresponding computation times were 4.4 and 0.2 CPU hours respectively. Using Beagle, stepwise imputation approaches improved accuracy when imputing to the smaller reference with 514 individuals, but showed no gain in accuracy when imputing to the reference with 11,457 individuals.

Keywords: imputation, accuracy, whole-genome sequence, stepwise, Beagle, FImpute

Introduction

Inclusion of well-chosen sequence variants and corresponding biological knowledge in statistical models shows promise for improving genomic prediction (Brøndum et al., 2015; MacLeod et al., 2016; O’Connell et al., 2016). Sequence variants of interest in the implementation of these approaches must be either genotyped directly or imputed using existing genotypes. The cost of genotyping all sequence variants of interest is high, but imputation of individuals genotyped on other panels is a cost-effective method for generation of these variants. Quantifying imputation accuracy and determining the best strategy for imputing variants of interest into the population are of key importance because incorrectly imputed genotypes can add noise and compromise prediction accuracy (Khatkar et al., 2012;
Chen et al., 2014). Studies have shown that imputation accuracy is dependent on a number of factors including reference size, relatedness between individuals and the imputation software used (Druet et al., 2010; Hickey et al., 2012; Ma et al., 2013). It has also been shown that stepwise approaches can be more accurate (Khatkar et al., 2012; Larmer et al., 2012; van Binsbergen, et al., 2014). The objectives of this study were to quantify accuracy and computation times for a number of different imputation strategies. Specifically, comparisons were made between imputing to different sequence references using two different imputation programs, Beagle and Flmpute. Direct vs stepwise imputation approaches via intermediate Illumina BovineSNP50k and BovineHD panels were also assessed.

Material and methods

Variants of interest

Selection of informative variants from whole-genome sequence (WGS) had previously been undertaken in another study. Variants were identified as informative according to the percentage variance they explained across a number of phenotypes for ~65k individuals within a BayesRC (MacLeod et al., 2016) framework. Further details of this variant identification approach are outlined elsewhere (Sherlock et al., 2018). Henceforth, this set of variants of interest will be referred to as the enriched set.

Imputation references and targets

Whole-genome sequence for 563 mixed-breed (Holstein-Friesian, Jersey and crossbreed) dairy cattle was aligned, called and filtered based on quality controls as described elsewhere (Sherlock et al., 2018). After filtering, 19,569,360 autosomal variants remained, which were subsetted to include only the enriched set of 158,238 variants. Validation individuals were identified as those which had WGS but had not previously been genotyped. For these 49 validation individuals, the enriched set was filtered to include only the set of 7,934 SNP in common with the latest GGP-LD panel. All other variants were masked.

Validation individuals were imputed to two different references:
1) Reference of non-validation individuals with WGS, across the enriched set of interest (514 individuals; 158,238 SNP).
2) Reference population of genotyped individuals with imputed sequence across the enriched set of interest (11,457 individuals; 158,238 SNP). Reference individuals were identified as those that had not been sequenced but had been genotyped on the Illumina BovineHD (n=145), BovineSNP50k (n=8,852) or custom versions of the GGP BeadChip (n=2,460), and were either males with progeny or their dams (8,994 males; 2,463 females). Imputed sequence for the reference individuals was generated by imputing their panel genotypes up to a sequence reference that included only the 514 non-validation individuals and 19,569,360 variants. Imputation was carried out using Beagle 4.0 (Browning & Browning, 2007), and the imputed sequence was subsetted to include only the enriched set of variants of interest.

Average pedigree relationships between the validation set and reference individuals were 0.029 (sd=0.039) and 0.026 (sd=0.031) for the two references respectively. The number of validation individuals with at least one parent in the reference was 26 and 17 for the two references respectively.

Imputation
Validation individuals were imputed to the two different references using direct and stepwise approaches. The direct approach involved imputing directly from a 7,934 variant set to the 158,238 variant set. The stepwise approaches involved imputing via a Bovine50k reference with 46,621 variants and 10,486 individuals and/or a Bovine HD reference with 675,321 variants and 3,334 individuals. The 50k reference had 5,389 variants in common with the 7,934 variants on the latest GGP-LD panel and 13,686 variants in common with the 158,238 enriched set. The HD reference had 7,562 variants in common with the 7,934 variants on the latest GGP-LD panel and 31,095 in common with the 158,238 enriched set. Imputation was performed using Beagle 4.0 (Browning & Browning, 2007) and FImpute (Sargolzaei et al., 2011) software with default parameters. The measure of accuracy used in this study was the correlation between observed and actual genotypes, and accuracy was summarised for the 150,304 variants in the enriched set but not on the latest GGP-LD panel.

**Results and Discussion**

Table 1 summarises imputation accuracy and computation times for both imputation references using direct and stepwise strategies. Across all strategies presented, imputing to a large reference of 11,457 imputed individuals was consistently more accurate than imputing to a smaller reference of 514 sequenced individuals. This corroborates evidence from other studies showing that an increase in reference size results in higher accuracy (Druet et al., 2010; Khatkar et al., 2012). Computation times using the large reference were also consistently higher. Stepwise strategies were also examined for FImpute but resulted in lower accuracies and higher CPU times compared to the direct approach and are not shown here. Stepwise approaches using Beagle improved accuracy when imputing to the reference of 514 individuals, with most improvement gained by imputing via the BovineHD panel only. However, stepwise approaches with Beagle showed no gain in accuracy when imputing to the reference of 11,457 individuals. This is not consistent with the findings of other studies such as Khatkar et al. (2012) and van Binsbergen et al. (2014) which found that accuracy was improved when using stepwise approaches. Imputation accuracy was higher for FImpute compared to Beagle when imputing to the smaller reference of true sequence, but accuracy was lower for FImpute when imputing to the large imputed sequence reference. Results from other studies also vary when comparing FImpute to Beagle with Ma et al. (2013) showing that Beagle had higher accuracy and Larmer et al. (2012) showing that accuracies using FImpute were similar or higher when imputing from 50k to HD. Overall, the fastest strategies were to impute directly to the reference of 514 true sequence individuals using either Beagle or FImpute (0.03 and 0.2 CPU hours respectively). Practically, computation for larger numbers of individuals may be required to fit within specific time constraints, so although these are not the most accurate strategies, they may still be attractive possibilities. In such instances, the trade-off between accuracy and computation time should be considered carefully.

**Conclusion**

Imputing to a large reference of 11,457 individuals that included imputed sequence variants was consistently more accurate than imputing to a smaller reference of only 514 individuals with true sequence. This demonstrated the value of including imputed variants in an imputation reference. Stepwise approaches did not consistently improve imputation accuracy,
and the most accurate strategy was to use Beagle software with a direct imputation approach. Other strategies requiring less computation time achieved lower but comparable levels of accuracy, thus still making them attractive options when there is a requirement to reduce computation time.

**List of References**


Table 1. Genotypic correlations and computation times for strategies using Beagle.

<table>
<thead>
<tr>
<th>Imputation reference</th>
<th>Beagle</th>
<th>F Impute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>Via 50k</td>
</tr>
<tr>
<td>True seq²</td>
<td>0.946</td>
<td>0.950</td>
</tr>
<tr>
<td>Imputed seq³</td>
<td>0.962</td>
<td>0.959</td>
</tr>
</tbody>
</table>

¹ Correlation between observed and actual genotypes
² Reference of 514 non-validation individuals and 158,238 variants
³ Reference of 11,457 individuals with imputed sequence across 158,238 variants