Imputation accuracy of whole-genome sequence data in Hanwoo cattle

Priscila A. Bernardes¹,⁵, Hawlader A. Al-Mamun¹, Mattias Suarez², Dajeong Lim³, Byoungho Park³ & Cedric Gondro¹,⁴

¹School of Environmental and Rural Science, University of New England, NSW, Australia
²Department of Primary Industries, NSW, Australia
³Animal Breeding and Genetics Division, NIAS, RDA, Republic of Korea
⁴College of Agriculture and Natural Resources, Michigan State University, MI, USA
⁵Departamento de Ciências Exatas, Universidade Estadual Paulista, SP, Brasil

Summary
This study reports the accuracy of imputation to whole genome sequence (WGS) from lower density genotype data in Korean Hanwoo cattle. A total of 203 reference Hanwoo cattle were used to impute 4,887 animals genotyped with Illumina BovineSNP50 v2 (50k). Also, a set of 928 animals genotyped with Illumina BovineHD chip (HD, 770k) were used as an intermediate reference dataset. First, the animals were imputed from 50k to HD and then imputed from HD to sequence. The accuracy of imputation was evaluated in animals in common between the reference and imputed population. Haplotype phasing and imputation was done using Eagle2 and Minimac3 software respectively. The imputation accuracy was calculated as the correlation between the imputed and observed genotypes and also as the percentage of correctly imputed genotypes (concordance rate). We achieved a mean imputation accuracy of 0.97 from 50k to WGS. The imputed data is useful to explore GWAS and genomic prediction for different commercially important traits in the Korean Hanwoo cattle.

Keywords: Hanwoo cattle, imputation accuracy, whole genome sequence data

Introduction
The development of new sequencing technologies and the rapid drop of sequencing costs have made whole genome sequencing viable for livestock species. However, sequencing a large number of animals, such as what is needed for genomic prediction, is still not routinely feasible (Brøndum et al., 2014). Therefore, researchers are still relying on imputation for genome wide association studies (GWAS) and for genomic predictions using whole genome sequence (WGS) data (MacLeod et al., 2016, VanRaden et al., 2017). The process of inferring unknown genotypes for the animals genotyped at lower density (i.e. 12k, 50k etc.) with the help of a set of reference animals genotyped at higher density (i.e. 700k, WGS, etc.) is called imputation. The use of imputed genotypes in genomic selection requires that all animals have their single nucleotide polymorphisms (SNPs) inferred with high accuracy (Badke et al., 2014). Highly accurate imputation gives us the opportunity to implement very cost effective genomic selection and gene discovery using less sequence data. In cattle, several studies report the accuracy of imputation from different low density panels to high density panels. Also there are many reports comparing imputation accuracies between different software developed for imputation. However, few
studies as yet have reported imputation accuracies using whole genome sequence data.

According to Van Binsbergen et al. (2014), the imputation accuracy can be affected by the size of the reference population, number of SNP in the panel(s), minor allele frequency (MAF) of SNPs to be imputed, the extent of linkage disequilibrium (LD) and population structure. In this study, we report the accuracy of imputation from lower density genotypes (50k) to whole genome sequence in Korean Hanwoo cattle. To our knowledge, this is the first study to impute low density Hanwoo cattle up to whole genome sequence, therefore it is important to evaluate imputation accuracy before using the imputed data for GWAS or genomic prediction.

Material and methods

Reference genotype data. Whole-genome sequence data was available for 203 Hanwoo cattle (main sires used to commercial population). The raw WGS data was pruned and aligned to the UMD3.1 bovine reference genome and variant calling was performed with a multi-sample variant calling pipeline that was implemented with the mpileup module of SAMtools (Li et al., 2009).

Target genotype data. A total of 4,887 animals genotyped with Illumina BovineSNP50 v2 (50k) and 928 animals genotyped with Illumina BovineHD chip (HD, 770k) were used as the target animals. These animals were from a half-sib family and average, minimum and maximum numbers of progeny/sire are 9, 0, and 35 respectively. Also there are 1783 animals in the target population having at least 1 parent in the reference population. Two steps were performed to impute the samples to WGS. First, all the animals with HD genotypes were used as the reference genotypes and all the animals with 50k genotypes were imputed up to 770k. There were 47,486 SNPs and 655 animals in common, presenting the HD and 50k genotypes. Next, the 4,887 imputed HD genotypes were imputed up to sequence level using 203 WGS genotypes as reference animals. The imputed HD and WGS presented 668,998 SNPs in common. There were 140 animals represented in both the 50k and WGS data. The total number of imputed variants in the reference set was 25,676,502, after exclusion of variants with minor allele counts of less than 4 (in each reference set) and with more than 2 allele types. Among these variants, 19,968,000 (77.76%) presented minor allele frequency (MAF) between the range 0.01<MAF<0.99.

Imputation. Imputation was done one chromosome at a time and before imputation both reference and target genotypes were phased per chromosome as well. Eagle version 2.3.2 (Loh et al., 2016) software was used for haplotype phasing and in this stage the missing genotypes in reference data were imputed. The imputation was carried out using Minimac3 software (Howie et al., 2012). The number of SNPs that passed to different thresholds for R² filter from the Minimac3 statistic was evaluated. The imputation accuracy was calculated as the correlation between the imputed and observed genotypes and the percentage of correctly imputed genotypes (concordance rate) using all the animals in common, which were present in the reference and the target genotypes. Therefore, the first step used 655 animals and 638,775 SNPs to calculate the imputation accuracy and the second used 140 animals and 25,007,504 SNPs.

Results and Discussion
The number of SNPs that passed based on different $R^2$ thresholds from the Minimac3 statistic is presented in Table 1. The average of imputation accuracy measured by the correlation between imputed 50k and reference HD was 0.97 when measured by animal and 0.85 when measured by SNP. Also, the average concordance rate was 97.34% by animal. From HD to sequence, the correlation ranged from 0.57 to 0.94 with an average of 0.92 when measured by animal and equal 0.87 when measured by SNP. Similar correlations were reported by Pausch et al. (2017) in dairy cattle. The concordance rate by animal ranged from 75.58% to 96.47% with an average of 94.61%.

Table 1. The number of SNPs that passed based on different $R^2$ thresholds

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Number of SNPs passed the filter</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2 &gt; 0.4$</td>
<td>15,310,805</td>
<td>59.63%</td>
</tr>
<tr>
<td>$R^2 &gt; 0.6$</td>
<td>12,612,785</td>
<td>49.12%</td>
</tr>
<tr>
<td>$R^2 &gt; 0.8$</td>
<td>8,890,931</td>
<td>34.63%</td>
</tr>
</tbody>
</table>

Only four animals had a correlation lower than 0.89 and a concordance rate of less than 92.51%. For these particular animals, more than 40% of the common (high MAF) SNPs were different, while for the other animals this was less than 5%. This could be a problem when the software estimated the haplotypes, which in turn resulted in wrong imputation of the SNPs. This could be avoided by applying a quality control to the panels. The correlation and concordance rate measured by SNPs are shown in Figure 1 for each chromosome.

![Figure 1](image_url)

Figure 1 – Boxplots and average (blue point) of imputation accuracy (correlation and concordance rate - %) by chromosome in Hanwoo cattle.

The presence of only 140 animals to verify the accuracy can affect the correlation estimates, which had lower averages compared to the concordance rate. There were also about 8 million SNPs for which it was not possible to calculate the correlation per SNP due to the absence of variability in either the imputed or the observed genotypes. In general, the chromosomes showed higher median than average values, which was influenced by some SNPs
with very low accuracy. We also observed variation in accuracy among different regions within the same chromosome. A large variation in imputation accuracy from HD to sequence, in different genomic regions, was also reported by Van Binsbergen et al. (2014) for dairy cattle; and the authors reported difficulties in obtaining a high accuracy of imputation in regions where the distances between SNPs on the HD panel was large. Pausch et al. (2017) reported high imputation errors in regions where the bovine genome presents segmental duplications. Possible differences between genotype calling in the 50k panel and WGS, regions with duplications, as well as the frequency and distance of SNPs to be imputed could be some of the reasons for getting lower imputation accuracy than expected.

In conclusion, we were able to achieve a mean imputation accuracy of 0.97 from 50k to WGS with the help of about 203 reference animals using Minimac3 with Eagle2 for pre-phasing. Although out of the more than 25 million imputed SNPs only about 15 million SNPs passed the filter threshold of $R^2 > 0.4$, we expect that the imputed data will be sufficient to explore GWAS and genomic prediction for different commercially important traits in Korean Hanwoo cattle.

**Acknowledgement**

CG and HAM were supported by a grant from the Next-Generation BioGreen 21 Program PJ01134906, and PJ012611 Rural Development Administration, Republic of Korea and Australian Research Council (DP130100542). P. A. Bernardes received scholarships from the São Paulo Research Foundation (Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP – process numbers 2016/22940-3 and 2015/25096-6).

**References**


Pausch, H., I.M. MacLeod, R. Fries, R. Emmerling, P.J. Bowman, H.D. Daetwyler, M.E.