J. Dairy Sci. 98:4131–4138
http://dx.doi.org/10.3168/jds.2014-9170

How imputation errors bias genomic predictions

E. C. G. Pimentel,[1] C. Edel, R. Emmerling, and K.-U. Götz
Institute of Animal Breeding, Bavarian State Research Center for Agriculture, Grub 85586, Germany

ABSTRACT

The objective of this study was to investigate in detail the biasing effects of imputation errors on genomic predictions. Direct genomic values (DGV) of 3,494 Brown Swiss selection candidates for 37 production and conformation traits were predicted using either their observed 50K genotypes or their 50K genotypes imputed from a mimicked 6K chip. Changes in DGV caused by imputation errors were shown to be systematic. The DGV of top animals were, on average, underestimated and that of bottom animals were, on average, overestimated when imputed genotypes were used instead of observed genotypes. This pattern might be explained by the fact that imputation algorithms will usually suggest the most frequent haplotype from the sample whenever a haplotype cannot be determined unambiguously. That was empirically shown to cause an advantage for the bottom animals and a disadvantage for the top animals.
Key words: allele frequency, bias, haplotype, single nucleotide polymorphism (SNP) effect

INTRODUCTION

In recent years, the number of genotyping platforms with different SNP densities has increased considerably. Additionally, custom chips containing any desired number of SNP defined by the customer are now commercially available. These increasing possibilities with respect to marker density make the role of imputation from one panel to another important. Many studies have been conducted on the effect of imputation on genomic predictions and their reliabilities, but results reported so far are usually given in terms of overall correlations between genomic predictions from observed and imputed genotypes (e.g., Dassonneville et al., 2011; Segelke et al., 2012). A closer inspection of the consequences of imputation errors on genomic predictions might be of interest. Therefore, the objective of this study was to analyze to what extent imputation errors affect genomic breeding values and to investigate whether the differences in predictions caused by imputation errors follow any systematic pattern.

MATERIALS AND METHODS

Brown Swiss data from the December 2013 run of the official German-Austrian joint genomic evaluation were used. The pool of genotyped animals included 3,494 selection candidates; that is, animals without insemination bull status and that do not contribute phenotypes to the system. Routine evaluations are based on the Illumina Bovine SNP50 BeadChip (Illumina Inc., San Diego, CA). After the usual edits (i.e., exclusion of markers with call-rate <0.95, minor allele frequency <0.02, significant deviation from Hardy-Weinberg equilibrium or redundancy with another locus), 37,653 markers remained for further analyses. Detailed descriptions of the major steps, the criteria used for marker editing, and the statistical method routinely used in the German-Austrian genomic evaluation can be found in Edel et al. (2011) and Ertl et al. (2014). In brief, the statistical model is

$$\mathbf{y} = \mu + \mathbf{Dg} + \mathbf{e},$$

where $\mathbf{y}$ is an $(n \times 1)$ vector of phenotypes of the calibration animals; that is, AI bulls contributing both genotypic and phenotypic information to the system; μ is an overall mean; $\mathbf{g}$ is a $(p \times 1)$ vector of direct genomic values (DGV), with $p = n + m$, and m being the number of selection candidates; $\mathbf{D}$ is an $(n \times p)$ design matrix relating phenotypes to DGV; and $\mathbf{e}$ is an $(n \times 1)$ vector of residuals. The variance of $\mathbf{y}$ ($\mathbf{V}$) is assumed to be

$$\mathbf{V} = \mathbf{DGD}'\sigma_a^2 + \mathbf{R},$$

where $\mathbf{G}$ is a $(p \times p)$ genomic relationship matrix, σ_a^2 is the additive genetic variance, and $\mathbf{R}$ is a diagonal matrix of order n, elements of which are functions of the residual variance and the reliability of the corresponding phenotype (for details, see Edel et al., 2009). Matrix

Received November 28, 2014.
Accepted February 20, 2015.
[1] Corresponding author: Eduardo.Pimentel@LfL.bayern.de

G was computed following the first method described by VanRaden (2008) as follows:

$$\mathbf{G} = \mathbf{ZZ}'\left[2\sum_{i=1}^{l} q_i\left(1-q_i\right)\right]^{-1},$$

where l is the number of markers; q_i is the base allele frequency at locus i; and $\mathbf{Z}$ is a ($p \times l$) matrix, calculated as $\mathbf{Z} = \mathbf{M} - \mathbf{Q}$, where $\mathbf{M}$ is the matrix of genotypes (coded as −1, 0, or 1) and $\mathbf{Q}$ is a matrix of which the ith column is $2(q_i - 0.5)$. Predicted DGV $(\hat{\mathbf{g}})$ are then calculated as

$$\hat{\mathbf{g}} = \hat{\sigma}_a^2\mathbf{GD}'\mathbf{V}^{-1}(\mathbf{y} - \hat{\mu}).$$

Reliabilities of DGV are estimated from

$$diag\left\{\mathbf{GD}'\mathbf{V}^{-1}\mathbf{DG}\right\}.$$

Phenotypes used in the analyses were deregressed multiple across-country evaluation (MACE) proofs. Estimates of base allele frequencies were obtained using the method proposed by Gengler et al. (2007). The DGV of the selection candidates for 37 production and conformation traits were predicted using either their observed 50K genotypes or their 50K genotypes imputed from a 6K chip. Genotypes of the 6K chip were obtained by masking the SNP from 50K that are not contained in the Illumina BovineLD BeadChip. Animals in the calibration set were all genotyped with the 50K chip. Masking of genotypes was only applied to selection candidates to depict a situation in which candidates are genotyped at low density. The number of calibration bulls varied depending on the trait and ranged from 1,001 to 5,390, with an average of 3,438. Imputation was done with 2 imputation software packages: findhap v2 (VanRaden et al., 2011) and FImpute (Sargolzaei et al., 2014). The number of animals with 50K genotypes in the reference population used for imputation was 6,243. From the 37,653 markers that passed the routine filtering process, 908 were not annotated. Therefore, these markers are meaningless for haplotype reconstruction and were not included in the imputation step done with findhap or FImpute. Genotypes at these loci were imputed with the sample mean gene contents (i.e., mean genotypes of the 6,243 reference animals with 50K genotypes) afterward.

RESULTS AND DISCUSSION

Average allele error rates, measured as the mean proportion of wrongly imputed alleles, were 1.54% with findhap and 0.85% with FImpute. Mean proportions of correctly imputed genotypes were 96.97% with findhap and 98.33% with FImpute. Mean correlation coefficients between observed and imputed genotypes were 0.976 with findhap and 0.987 with FImpute. These numbers are similar to measures of imputation success from 6K to 50K reported in other studies (e.g., Boichard et al., 2012; Segelke et al., 2012; Chen et al., 2014). Across traits, average overall correlations between DGV predicted with observed or imputed genotypes were 0.987 (from 0.982 to 0.993) with findhap and 0.992 (from 0.988 to 0.995) with FImpute. These numbers are similar to correlations reported in other studies (e.g., Mulder et al., 2012; Segelke et al., 2012). Despite these overall high correlations, some noticeable reranking among the top selection candidates occurred when prediction was based on imputed genotypes. Averaged across all traits, rank correlations within the top 50 candidates were 0.843 with findhap and 0.876 with FImpute. Within the top 50 candidates, we found a tendency to underestimation when DGV were predicted from imputed genotypes. Analogously, a tendency to overestimation within the bottom 50 candidates was observed. As an illustration, mean differences between DGV from observed genotypes and from genotypes imputed with findhap for the bottom 50, intermediate, and top 50 candidates (ranked according to the DVG from observed genotypes) are given in Figure 1 for 6 of the studied traits. These trends indicate that the changes in DGV caused by imputation errors follow some systematic pattern. As a possible explanation to this phenomenon, we formulated a hypothesis based on the following 3 assumptions: (1) in a simplified way, one could postulate that the top animals should have, on average, the best haplotypes, and that the bottom animals should have, on average, the worst haplotypes, with respect to their effects on the trait being considered; (2) whenever an imputation algorithm cannot determine a haplotype unambiguously, it will suggest the most frequent haplotype in the sample as replacement for the missing one; and (3) if the most frequent haplotype has a neutral effect on the trait (i.e., if its effect is the closest to the population mean compared with the effects of the other possible haplotypes), then this replacement will represent an advantage for the bottom animals and a disadvantage for the top animals.

For most of the traits, we observed a general decrease in DGV when imputed genotypes were used. This trend can be seen in Figure 1 as the slight decrease in DGV for the intermediate animals. This overall decrease can be attributed to the genetic trend that separates the group of selection candidates from the calibration group (and the reference pool of genotyped animals used for imputation). Compared with the reference group, selection