How do Imputation Errors Affect Genomic Breeding Values?

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

Abstract

The objective of this study was to investigate in more detail the biasing effects of imputation errors on genomic predictions. Genomic breeding values (**GEBV**) of 3494 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 GEBV caused by imputation errors were shown to be systematic. The GEBV of top animals were on average underestimated and GEBV 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, SNP effect

Introduction

In recent years the number of genotyping platforms with different single nucleotide polymorphism (**SNP**) densities has considerably increased. Additionally, customized 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 platform to another important. Many studies about the impact of imputation on genomic predictions and their reliabilities have already been done, 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 therefore be of interest.

The objective of this study was to analyze to which extent imputation errors affect genomic breeding values and to closer investigate whether the differences in predictions caused by imputation errors follow any systematic pattern.

Material 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 3494 selection candidates, i.e., animals without insemination bull status. Routine evaluations are based on the Illumina Bovine SNP50 BeadChip. After the usual edits, 37653 markers remained for further analyses. Detailed descriptions of the criteria for marker editing and the statistical method routinely used in the German/Austrian genomic evaluation can be found in Edel *et al*. (2011) or Ertl *et al.* (2014). Genomic breeding values (**GEBV**) 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 SNPs from 50k that are not contained in the Illumina BovineLD BeadChip. Imputation was done with two imputation software packages: findhap v2 (VanRaden *et al.*, 2011) and FImpute (Sargolzaei *et al.*, 2011).

Results and Discussion

Across traits, average overall correlations between GEBV predicted with observed or with imputed genotypes were 0.988 (from 0.983 to 0.993) with findhap and 0.992 (from 0.989 to 0.995) with FImpute. Despite these overall high correlations, some noticeable re-ranking among the top bulls occurred when prediction was based on imputed genotypes. Some of the changes within the top 50 candidates are illustrated in Table 1 for nine of the studied traits. Averaged across all traits, rank correlations within the top 50 list were considerably lower than the correlations across all animals. Classification of the candidates as belonging to the top 50 also differed when ranking was based on GEBV predicted from observed or from imputed genotypes.

As indicated by the correlations across all candidates, there was an overall good agreement between GEBV predicted from observed and from imputed genotypes. Within the top 50 candidates, there was a tendency to underestimation when GEBV were predicted from imputed genotypes. Analogously, a tendency to overestimation within the bottom 50 candidates could be observed. These patterns are depicted in Figures 1 and 2 for Protein (kg) as an illustration. Similar tendencies were also observed for all the 37 traits analyzed. This gives an indication that the changes in GEBV caused by imputation errors follow some systematic pattern. As a possible explanation to this phenomenon, we formulated a hypothesis based on the following three 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 to the missing one; (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, in comparison to 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.

The analyzed data were used to calculate some statistics in order to investigate if the above mentioned assumptions may hold. The first assumption does not need to be addressed, because the model used for predicting GEBV here implies exactly what was formulated in point (1). The second assumption is actually in agreement with the descriptions of the algorithms used in population imputation (e.g., VanRaden *et al.*, 2011). Nevertheless, we checked in the imputed genotype data set how often an incorrectly imputed allele was the most frequent one at its locus. For each trait, software (findhap and FImpute) and group (top and bottom), the mean proportions of changes for the most frequent allele are given in Figure 3. Changes for the most frequent allele occurred on average more often than changes for the least frequent allele in almost all cases, which is in agreement with the statement in assumption (2). Differences were more evident when imputation was performed with FImpute, whilst differences with findhap were small but consistent across traits and groups. Quantitative genetics theory shows that the higher the frequency of an allele, the smaller is the deviation of its effect from the population mean. To investigate if the third assumption might hold, we looked at the SNP effects on each of the analyzed traits and checked whether an incorrectly imputed allele had an increasing or decreasing effect on the breeding value. For each trait, software and group, the mean proportions of changes for an allele with a positive or with a negative effect on the trait were computed. Results are given in Figure 4. One can see that within the group of top candidates changes for an allele with a negative effect on the trait occurred more often than changes for an allele with a positive effect. Analogously, within the group of bottom candidates changes occurred more often to an allele with a positive effect. These patterns were similar for both software and are also in agreement with the assumption made in point (3). The statistics calculated from the analyzed data are in good agreement with the expectations if the assumptions made above were correct. This does not prove the formulated hypothesis, but gives strong empirical evidence that it may hold.

Conclusions

Imputation errors seem to cause systematic changes in genomic predictions, which tend to be underestimated in the top segment and overestimated in the bottom segment. This pattern might be explained by the fact that imputation algorithms will usually suggest the most frequent haplotype observed in the sample as replacement to the missing one whenever a haplotype cannot be determined unambiguously. This feature of imputation was empirically shown to induce an advantage to animals in the bottom and a disadvantage to animals in the top segment. That might have implications in genomic evaluations, especially with data pools comprising animals genotyped at different densities and strong selection. In such cases, good selection candidates genotyped at low density panels could be penalized.

Acknowledgements

We thank Paul VanRaden for making the software findhap freely available online, Mehdi Sargolzaei for kindly providing a trial version of FImpute and Intergenomics for many of the genotypes used.

References

Dassonneville, R., Brøndum, R.F., Druet, T., Fritz, S., Guillaume, F., Guldbrandtsen, B., Lund, M.S., Ducrocq, V. & Su, G. 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.

Ertl, J., Edel, C., Emmerling, R., Pausch, H., Fries, R. & Götz, K.-U. 2014. On the limited increase in validation reliability using high-density genotypes in genomic best linear unbiased prediction: Observations from Fleckvieh cattle. *J. Dairy Sci. 97*, 487–496.

Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S. 2011. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics 15:478*.

Segelke, D., Chen, J., Liu, Z., Reinhardt, F., Thaller, G. & Reents, R. 2012. Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. *J. Dairy Sci. 95*, 5403–5411.

VanRaden, P.M., O'Connell, J.R., Wiggans, G.R. & Weigel, K.A. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol. 43*, 10.

Table 1. Changes in ranking within the top 50 candidates when predicting genomic breeding values (GEBV) with observed 50k genotypes or with 50k genotypes imputed from a 6k chip.

Trait	Rank correlation[a]		Also top 50 in imputed set[b]	
	findhap	FImpute	findhap	FImpute
Milk (kg)	0.82	0.90	42	44
Fat (kg)	0.90	0.91	42	46
Protein (kg)	0.82	0.91	42	43
SCS	0.79	0.87	43	41
Workability	0.71	0.88	40	44
Udder depth	0.89	0.89	42	40
Feet & legs	0.89	0.93	45	44
Udder	0.80	0.84	44	42
Overall score	0.86	0.89	44	43
Average (37 traits)	**0.84**	**0.88**	**44**	**43**

[a]Spearman correlation coefficient between GEBV from observed 50k and from imputed genotypes, of the top 50 candidates ranked according to GEBV from observed 50k genotypes.

[b]Number of bulls from a top 50 list, ranked according to GEBV from observed 50k genotypes, that are also present in a top 50 list ranked according to GEBV from imputed genotypes.

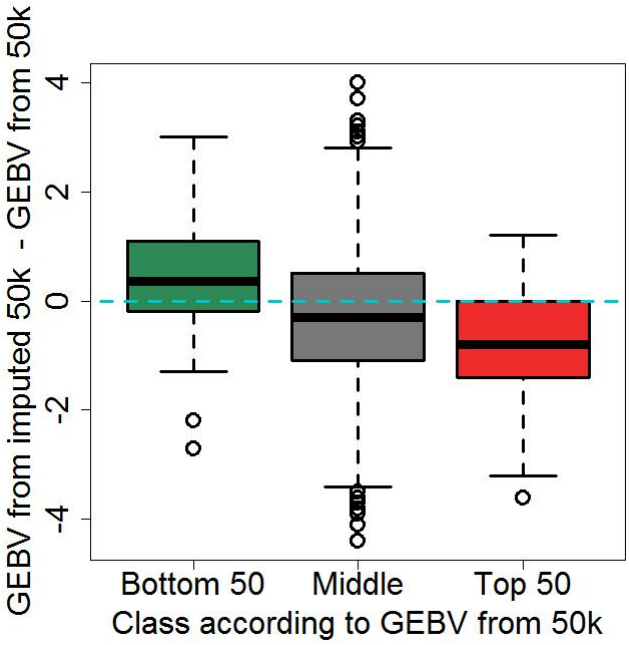

Figure 1. Differences between genomic breeding values (GEBV) for Protein (kg) predicted from imputed (with FImpute) and from observed genotypes for classes of top, bottom and intermediate candidates ranked according to their GEBV from observed 50k.

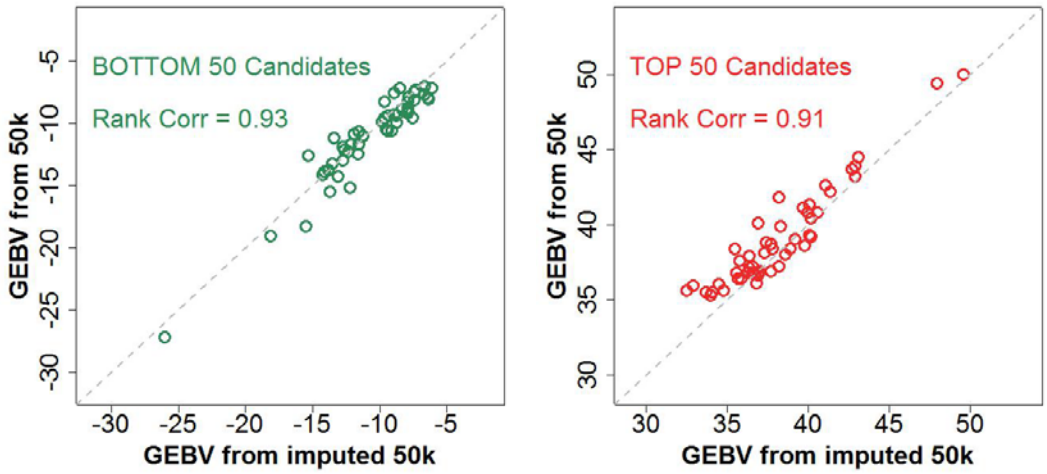

Figure 2. Genomic breeding values (GEBV) for Protein (kg) predicted from observed genotypes against GEBV predicted from imputed genotypes (with FImpute) within the groups of bottom and top candidates.

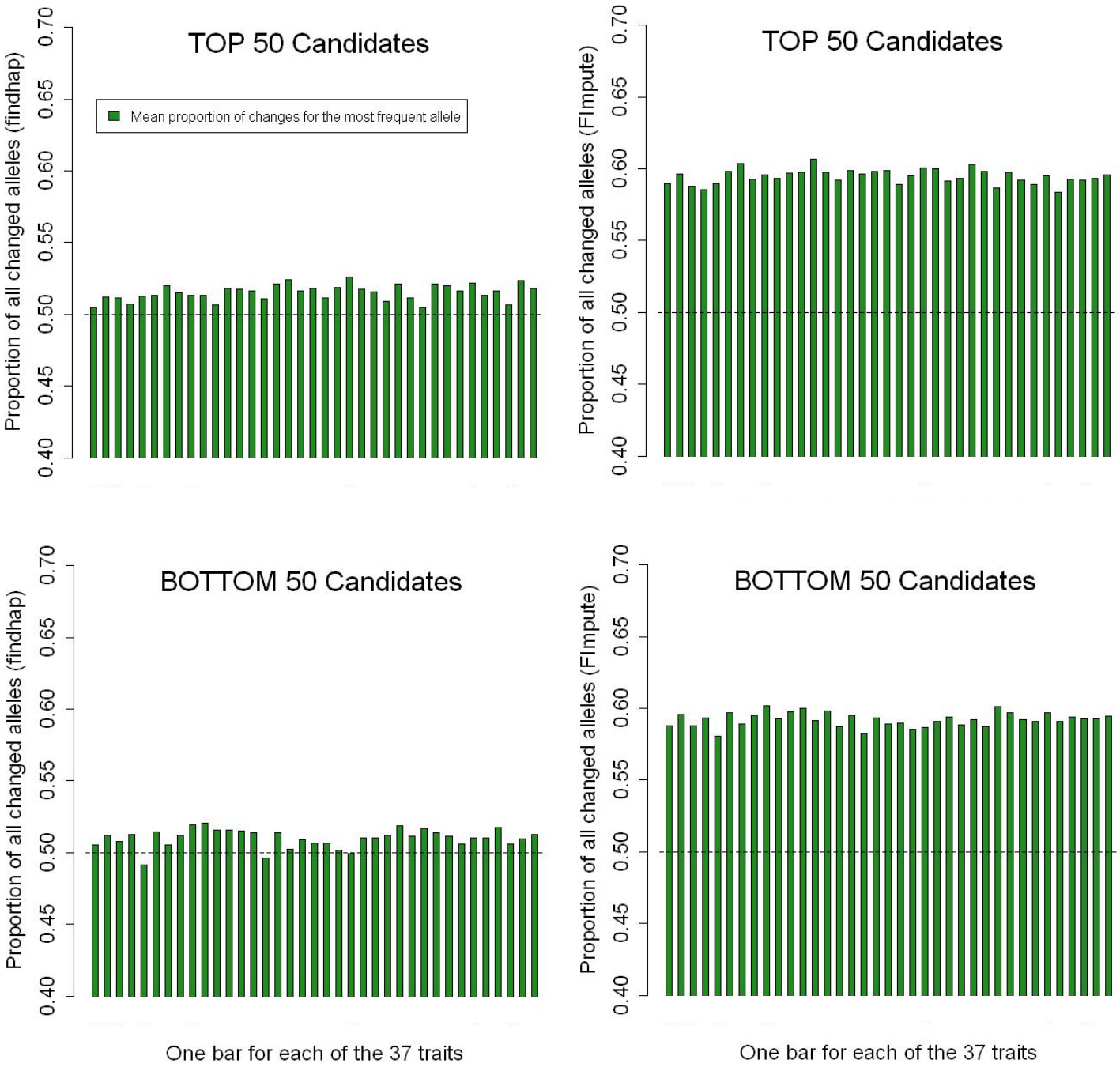

Figure 3. Mean proportion of the incorrectly imputed alleles that were the most frequent of its locus, within the top and bottom candidates, and imputed with findhap or FImpute.

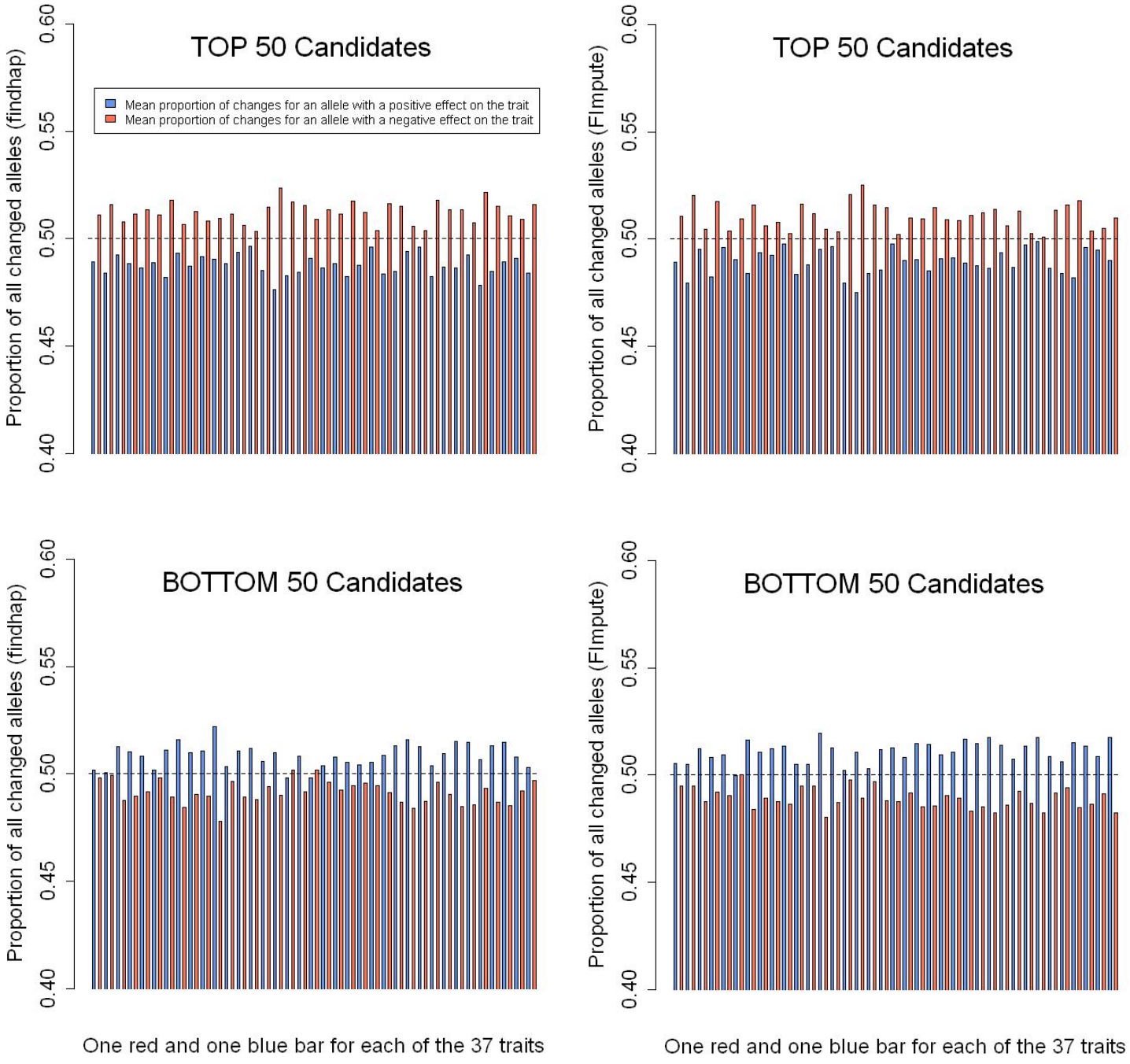

Figure 4. Mean proportions of the incorrectly imputed alleles that were the ones with a positive (blue bars) or with a negative (red bars) effect on the trait, within the top and bottom candidates, and imputed with findhap or FImpute.