Reliability of Genomic Prediction Using Imputed Genotypes for German Holsteins: Illumina 3K to 54K Bovine Chip

J. Chen, Z. Liu, F. Reinhardt and R. Reents
***vit** w.V., Heideweg 1, 27283 Verden, Germany*

Abstract

Low-cost low-density SNP chips are developed for enabling large-scale genotyping of animals at a reasonable accuracy of genomic prediction. In order to assess the accuracy of a low density chip Illumina Bovine3K chip, genotypes of animals of the 3K chip were simulated using the current Illumina Bovine50K information. Three imputing softwares, Beagle, DAGPHASE and Findhap, were applied to three genotype data sets: German Holstein bulls, EuroGenomics Holstein bulls, and all genotyped animals of German Holstein breed. The imputed 54K genotypes were used to calculate DGV and combined GEBV following routine procedures of genomic evaluation for German Holsteins (Liu *et al.*, 2011). To evaluate imputing accuracy of the three softwares, 1369 youngest German Holstein bulls, born between September 2003 and December 2004, were chosen as validation animals. The three imputing softwares differed in computing time markedly, Findhap being much faster than Beagle and DAGPHASE. Allele error rate for the EuroGenomics bull dataset was 3.3% for Findhap, 2.7% for DAGPHASE, and 1.6% for Beagle, respectively. Phenotypic data from April 2010 Interbull evaluation were used to assess the loss in accuracy of genomic prediction using the imputed 54K genotypes of EuroGenomics data set. Equal regression coefficients, by regressing deregressed EBV of the validation bulls, were obtained with the imputed 54K genotypes as the real ones, indicating that GEBV of the imputed 54K genotypes were as unbiased as using the real genotypes. However, R^2 value of GEBV of the imputed genotypes decreased by: 5.0% for Findhap and 2.1% for Beagle, across all the evaluated traits. On average, reliability of GEBV dropped by 6.5% for Findhap and 2.6% for Beagle, respectively. Based on the differences in computational requirements and imputing accuracy, different imputing softwares may be chosen for large-scale routine genotype imputation and genomic evaluation or for small-scale imputation without time constraints.

1. Introduction

Currently, routine genomic evaluation for German Holstein (Liu *et al.,* 2011) uses Illumina Bovine50K chip. With the introduction of the low density Illumina Bovine3K chip, dairy cattle breeders are given a new opportunity for genotyping animals at a lower cost. On the other side new chips of higher density, or even complete genome sequencing, are available now for dairy cattle breeding. Dairy geneticists are challenged to work with ever more diverse SNP chips (VanRaden *et al.,* 2011). The objectives of this study were 1) to assess the accuracy of imputed 54K genotypes from 3K chip, and 2) to quantify the loss in reliability of genomic prediction using the imputed genotypes.

2. Materials and Methods

Original Illumina Bovine50K genotypes were obtained from **vit**'s routine genomic evaluation in February 2011 for the 3K imputing study. A total of 34,802 genotyped Holstein animals were selected and low-density 3K genotypes of the animals were simulated according to the 3K marker map from Illumina (Illumina Inc., San Diego, CA). All 54001 SNP markers from the Bovin50K chip were considered, not only those (n=45,181) used in genomic evaluation for German Holsteins (Liu *et al.,* 2011). A total of 1672 SNPs were discarded since their chromosomal locations were not known. Three imputing softwares were applied to the 3K genotypes: DAGPHASE (Druet and Georges, 2010 version 2.3), Beagle (Browning and

Browning 2010, version 3.3), and Findhap (VanRaden *et al.*, 2011, version 2). Three data sets (Table 1) were used for the investigation of imputing accuracy of the three softwares: German national genomic reference population (data I), EuroGenomics reference population (data II), and all genotyped Holstein animals (data III). As validation animals, 1369 German Holstein bulls born between September 2003 and December 2004 or genotyped animals born after June 30th, 2010, were used for data sets I and II or III, respectively.

Table 1. Genotype data for the 3K imputation.

	Reference population	Validation population
Data I: German domestic bulls	3589 (born before Sep 2003)	1369 (born Sep 2003 ~ Dec 2004)
Data II: EuroGenomics bulls	14,385 EuroGenomics Holstein bulls (born before Sep 2003)	1369 DEU bulls
		1457 DFS, FRA, NLD bulls
Data III: all genotyped Holstein animals	32,597 animals born before July 2010	2205 animals born >= July 2010

Imputed 54K SNP genotypes were compared to their original ones for the validation animals of the three data sets. Allele error rates were calculated for every SNP, chromosome, and the whole genome as well as for every validation animal. The imputed 54K genotypes were further used for investigating the loss of accuracy of genomic prediction only for the data II with EuroGenomics bulls, because a genomic validation based on original 54K genotypes had been performed only for this scenario (Liu *et al.*, 2011).

3. Results

3.1. Accuracy of the 3K imputing

All calculations were conducted on a cluster of 64-bit Linux servers with multiple AMD Opteron processors each. Table 2 shows the computational requirements for the 3K imputations. Total computing times from multiple processors were summed up in hours for each of the three data sets.

Table 2. Computing requirements of the softwares for the 3K imputations.

Findhap DAGPHASE Beagle	Total computing time (hours)	Maximum RAM usage (Gb)
Data I:	0.8	1.4
German	244	0.4
domestic bulls	264	1.0
Data II:	5.6	3.7
EuroGenomics	1608	0.5
bulls	3960	3.0
Data III: All	14.3	5.6
genotyped	N/A	N/A
animals	N/A	N/A

It can be seen from Table 2 that the three imputing softwares differed markedly in computing time, with a clear advantage for Findhap. Due to too high computational demands of Beagle and DAGPHASE, we used only Findhap for the 3K imputing of the third data with 32,597 animals in reference population. Table 3 shows allele error rates of the 3K imputing for the three data sets. Validation bulls/animals were defined as Black-and-White Holstein bulls (data I & II) /animals (data III) with sire in the imputing reference population.

Table 3. Allele error rates of the 3K imputing.

Findhap DAGPHASE Beagle	Number of validation animals	Allele error rate (%)
Data I: German		2.6
domestic bulls	458	5.0
		1.7
Data II:		3.3
EuroGenomics	1019	2.7
bulls		1.6
Data III: All		2.7
genotyped	1881	N/A
animals		N/A

As a result of more bulls in the imputation reference population, allele error rate decreased from German national to EuroGenomics reference bull population, except DAGPHASE. It can be seen in Table 3 that Beagle gave the lowest and Findhap the highest error rate for all three data sets. Due to many more reference animals in data III, error rate of Findhap dropped to 2.7% in comparison to 3.3% of data II. Non-German EuroGenomics validation bulls from France, Nordic countries and The Netherlands had equal error rate as the German validation bulls in data II. When sire of a validation animal was not present in the imputing reference population, error rate increased by 1.5%. Error rate was 0.1% lower, if maternal grandsire of a validation animal belonged to the reference population. Since Red-and-White Holstein bulls accounted for only 10% reference bulls, their error rate was 1% higher than Black and White Holstein validation bulls. As reference bulls from The Netherlands had imputed Bovine54K genotypes from a customised 60K chip, progeny of the Dutch bulls had slightly higher error rate in the imputed 54K genotypes.

In the genomic validation study with original 54K genotypes late measured traits, such as longevity, had smaller reference population and older validation bulls than regular traits like milk yield. Therefore, a separate imputing validation was done with 11,737 EuroGenomics bulls in reference population born before 2002. Imputing error rate was 0.45% higher for the late measured traits than for regular traits with larger reference population for data II.

Table 4 shows observed correlations between direct genomic values (DGV) of validation bulls of EuroGenomics bull data (data II) for milk yield. The DGV correlation between Beagle and Findhap was 0.94. DGV of Beagle had higher correlation with deregressed EBV (DRP) than DGV of Findhap. We can also see difference in correlations of DGV with DRP or EBV between both softwares.

Table 4. Observed correlations of DGV for validation bulls of EuroGenomics data (data II).

	B	C	D	E
Imputing Findhap (A)	.94	.94	.69	.71
Imputing Beagle (B)		.96	.73	.74
Real 54K genotypes (C)			.74	.75
Deregressed EBV (D)				.99
Conventional EBV (E)				

3.2. Accuracy of genomic prediction using the imputed 54K genotypes

Table 5 shows R^2 value of regressing DRP of validation bulls on their GEBV using the imputed 54K genotypes. Compared to R^2 value increase by genomics, R^2 of GEBV – R^2 of pedigree index (PI), the imputed 54K genotypes had a reduction in the R^2 value of 5.0% for Findhap or 2.1% for Beagle for the selected nine traits, on average. Traits influenced by a major gene, such as fat and milk yield, had more decrease in R^2 value. In addition, R^2 value dropped more for traits with higher gain in genomic reliability, like somatic cell scores (SCS), stature, and udder depth, than those with lower genomic reliability gain, e.g. days open or body conditional score (BCS).

Table 5. Reduction in R^2 value of genomic prediction using imputed 54K genotypes.

R^2_{GEBV} – R^2_{PI}, %	Using real 54K genotype	R^2 drop (Findhap)	R^2 drop (Beagle)
Milk, kg	29.0	5.5	0.9
Fat, kg	28.4	6.4	3.1
Protein, kg	21.1	5.7	1.1
SCS	31.2	5.8	2.9
Days open	11.6	1.6	0.0
Longevity	22.8	5.1	1.0
Stature	32.5	6.4	3.4
U. depth	36.8	6.3	4.9
BCS	22.4	2.4	1.5
Average	26.2	5.0	2.1

Table 6 gives reduction in reliability of genomic prediction using the imputed 54K genotypes.

Table 6. Reduction in reliabilities of genomic prediction using imputed 54K genotypes.

(R^2_{GEBV} – R^2_{PI})/REL, %	Using real 54K genotype	Reliability drop (Findhap)	Reliability drop (Beagle)
Milk, kg	35.1	6.6	1.1
Fat, kg	34.6	7.8	3.8
Protein, kg	25.8	7.0	1.3
SCS	43.6	8.2	4.1
Days open	18.1	2.6	0.0
Longevity	35.2	7.8	1.5
Stature	37.0	7.3	3.8
U. Depth	46.9	8.0	6.3
BCS	28.5	3.1	1.9
Average	33.9	6.5	2.6

The R^2 value increase by genomics, R^2_{GEBV} – R^2_{PI}, was divided by average reliability (REL) of conventional EBV of the validation bulls. On average, reliability of GEBV decreased by 6.5% for Findhap and 2.6% for Beagle, respectively, for the selected traits.

Regression coefficients of deregressed EBV on GEBV were equal in the genomic validation using the imputed genotypes as using the real 54K genotypes. This indicates that using the imputed genotypes did not lead to biased GEBV. However, reliabilities of GEBV based on the imputed genotypes decreased. In reality, the reduction in genomic reliability due to the use of the imputed 54K genotypes may not be as high as in this study, because low-density 3K candidates tend to have more relatives genotyped with the 54K chip and the reference population of the 54K chip is significantly larger.

Effects of SNP of the original 3K chip, without being imputed to 54K, were estimated in a special test run. Observed correlation of DGV of candidates without phenotypes was e.g. 0.81 for milk yield between the genotypes of 3K and 54K. Adding pedigree index led to a slightly higher correlation of GEBV for the candidate, 0.85 for milk yield. Additionally, variances of DGV and GEBV of the candidates were significantly lower than those from 54K genotypes. It demonstrated clearly that an imputation of the 3K to 54K was necessary for a reasonably accurate genomic prediction.

4. Discussion

Using three softwares genotypes of the low-density Illumina Bovine3K were imputed to 54K for three different genotype data sets. The softwares differed significantly in computing time and varied in accuracy of imputation, with 1.6% allele error rate for Beagle, 2.7% for DAGPHASE and 3.3% for Findhap, respectively. GEBV using the imputed genotypes were as unbiased as using real 54K genotypes. However, reliabilities of GEBV based on the imputed genotypes decreased by 6.5% (Findhap) or 2.5% (Beagle), on average, for a group of selected traits. For large-scale routine genotype imputation and genomic evaluation efficiency of imputing software is more important than for small-scale imputing with no time constraints, due to the significant difference in computational efficiency between the imputing softwares.

5. References

Browning, B.L. & Browning, S.R. 2010. High-resolution detection of identity by descent in unrelated individuals. *Amer. J. Hum. Genet. 86,* 526-539.

Druet, T. & Georges, M. 2010. A hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics 184,* 789-798.

Liu, Z., Seefried, F., Reinhardt, F., Rensing, S., Thaller, G. & Reents, R. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genet Sel. Evol. 43,* 19.

VanRaden, P.M., O'Connell, J.R., Wiggans, G.R. & Weigel, K.A. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol. 43,* 9.