Impact of the Structure of the Genomic Matrix on Accuracy of Evaluations for Imputed Genotypes

Raphael Mrode*[1] and Mike Coffey*

*Scotland Rural College, Roslin Institute Building, Easter Bush, EH25, 9RG and [1]International Livestock Research Institute, Nairobi, Kenya

Abstract

Genotypes corresponding to 20 different low density chips currently imputed to the 50K illumina chip in the United Kingdom (UK) were created from bulls genotyped with the 50K chip. For each low chip type, 1200 bulls with genotypes were generated and then imputed to 50K. The accuracy of imputation was then computed as the correlation between the imputed genotypes and the original 50K genotypes within 4 classes of bulls: those with high, medium, low and no relationship with bulls in the reference (REF) population consisting of 14,280 bulls. Genomic relationship (**G**) matrices were computed using both the REF bulls and bulls with imputed genotypes. GBLUP was therefore undertaken with the bulls with imputed genotypes regarded as validation bulls for production traits and somatic cell counts. Four different types of **G** were used in the GBLUP: computed from genotypes but with 0.01 added to diagonal elements or (ii) with a weight of 0.05, 0.10 or 0.20 given to the **A** matrix. The impacts of these different G matrices on the genomic accuracies for the validation animals with imputed genotypes were evaluated. In general, the accuracy of imputation increased as validation animals have more relationship with animals in the REF population with accuracy varying from 0.87 to 0.999. Similarly, the accuracy of prediction increased as more polygenic effects were included in the construction of the **G** matrix for traits analysed. This increase was about 2 to 4 % for milk yield and 1 to 2% for SCC for most low density chips but was highest for chip 3. This could be attributed to more information coming from the A matrix resulting in increased accuracy.

Key words: imputation, accuracy, genomic prediction

Introduction

Currently in the UK, about 20 different low density chips are imputed to the 50K Illumina chip for over 115,000 bulls included in the national genomic prediction system. These low density chips vary in size from 3k to 80K. A description of the chips is shown in Table 1. Imputation of the low density genotypes to 50K is implemented using the software Findhap (VanRaden, 2010). The overall aim of the study was to understand the impact of imputing chips of varying densities on the accuracy of the gEBVs for bulls with the imputed genotypes. However, the accuracy of imputation for these bulls was initially assessed within four classes of bulls for each chip type: those with high, medium, low and no relationship with bulls in the reference population. Subsequently, the accuracy of gEBVs for bulls with imputed genotypes was computed within the same four categories. The usefulness of including a polygenic effect in genomic models to capture the proportion of genetic variance not captured by markers has

been demonstrated (Liu *et al.,* 2011). The question posed here is; does the inclusion of a polygenic effect in the model provides an additional benefit in terms of accuracy of gEBVs for bulls with imputed genotypes which have low or no relationship with bulls in the reference population or low relationships with all other bulls in the pedigree used for the imputation? This study therefore examined the impact that various genomic matrices constructed with different weights on the **A** matrix had on accuracies of genomic predictions within the 4 classes of bulls.

Materials and Methods

Genotypes datasets corresponding to SNPs found on 20 different low density chips currently in use in the UK were created from bull genotypes from 50K Illumina chips using their chromosome map. For each low density chip type, 1200 bulls with genotypes were generated and then imputed to 50K. Note that not all 50K SNPs are present on the higher

density chips, and therefore imputation of these is required. The accuracy of imputation was then computed as the correlation between the imputed genotype and the original 50K genotype within 4 classes of bulls: those with high, medium, low and no relationship with bulls in the reference population consisting of 14,280 bulls. Genomic relationship matrices were computed using both the reference bulls with actual genotypes and bulls with imputed genotypes. GBLUP was then undertaken with the bulls with imputed genotypes regarded as validation bulls for production traits and somatic cell counts (SCC), with SCC representing a trait with a lower heritability. The input variables for the analyses of these traits were de-regressed breeding values (DRPs). The accuracy of genomic prediction was computed as the correlation between direct genomic breeding values from GBLUP and the DRPs for the validation animals. The relative accuracy of using imputed genotypes was computed as the ratio of accuracy of prediction with imputed genotypes to that with the original genotypes for the validation animals. To evaluate the impact of the including polygenic effects on the relative accuracy of imputed genotypes, four different types of **G** were used in the GBLUP: computed from genotypes but with (i) 0.01 added to diagonal elements or (ii) with a weight of 0.05, 0.10 or 0.20 given to the **A** matrix.

Results and Discussions

Accuracy of imputation

The accuracy of imputation increased generally with the degree of relatedness with animals in the references population especially in the chips of the smallest size (Table 2). The maximum difference in accuracy was about 2.5% for the 3k chip between animals with an average relationship of 80% with the reference population and those with none. However, that difference is an increase of about 10% for some animals which in addition are poorly related to all animals in the pedigree used for the imputation (Table 5). The same pattern is generally observed for medium density chips with sizes varying from 17k to 30k in terms of

relatedness to the reference population (Table 3) but the maximum difference in accuracy is now less than 1% for animals related at about 80% with the reference population and those not related to the reference population. However, when chip sizes are greater than 30K, accuracy of imputation was essentially 0.999 for all categories of bulls; therefore results have not been presented.

Accuracy of genomic prediction

The accuracy of genomic predictions for the 4 classes of bulls is only presented for two low density (chips 3 and 6) and medium density chip 9 for milk yield and SCC. The results for fat and protein yields followed the same pattern as milk yield and have therefore not been presented. Firstly, considering evaluations based on actual genotypes, the accuracy of gEBVs for milk yield and SCC tended to be higher in the 4 classes of bulls as the average relationship of these bulls with the reference population increased (see footnotes of Figures 1 to 6). This trend is influenced at times by the fewer number in the first category of bulls which had the highest relationship with the reference population. With the use of imputed genotypes, the accuracy of gEBVs of bulls with least relationship with the reference population, increased as more polygenic effects were included in the construction of the **G** matrix (Figures 1 to 6). This increase was about 2% for milk yield and 1 to 2% for SCC for most low density chips but was highest for chip 3 showing an increase of about 4 % for milk yield as level of polygenic effects increased from 0 to 20%. This could be attributed to more information coming from the **A** matrix resulting in increased accuracy. Comparison of the **G** from actual genotypes with the imputed genotypes showed a slight reduction in the average relationship among bulls in categories 3 and 4 which were least related to the reference population. However as the percentage of **A** included in **G** computed using imputed genotypes increases, the average relationship among these animals increased as well accompanied by a reduction in the standard deviation of their average relationships.

Conclusion

In general for chips of low density the accuracy of imputation is influenced by the degree of their relatedness to bulls in the reference population. The accuracy increases as the relatedness increases. However, this tendency decreases in medium sized chips and is non-existent in chips more than 30K. For production traits and SCC, the accuracy of genomic predictions similarly increased as bulls are more related to the reference population. The inclusion of some degree of polygenic effects contributed to improved accuracy in the bulls poorly related to the reference bulls

Acknowledgement

References

Liu, Z., Seefried, F.R., 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. *Genetic Selection Evolution 43:*19.

VanRaden, P.M. 2010. Genomic evaluations with many more genotypes and phenotypes. Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany, Aug. 1–6, Comm. 27.

Table 1. A Summary and classification of current bulls imputed in the UK.

Low density chips of size 3k to 15k			Medium density chips of size 17k to 30k		
Chip number	Size	Name	Chip number	Size	Name
3	3000	3K	9	19725	GGP-Super
6	6909	LD	14	26151	GGP V3
7	8762	GGP	100	19720	GGP V1-1
10	11410	ZLD	109	19809	GGP-S+90
12	9072	ELD	110	20074	GGP_SuperLD_ SemiPrivate
13	6912	LD2	111	20077	GGP_SuperLDv1-1_SemiPrivate
16	14376	IDBV2	112	26359	GGPv3_SemiPrivate_E
			215	17619	ZL2
High density chips of size greater 30k					
8	76999	GHD			
11	56955	ZMD			
15	76883	GGPHD_T_SemiPrivate			
108	77068	GHD +68			
208	76934	GHD T			

Table 2. Accuracy of imputation for low density chips for 4 classes of bulls with different numerator average relationships (AVgRel) with bulls in the reference population.

	Low density chips if size 3k to 15k						
AVgRel	3	6	7	10	12	13	16
>80	0.922	0.978	0.981	0.986	0.982	0.979	0.988
25 – 80	0.917	0.974	0.979	0.985	0.981	0.976	0.986
1 - 24	0.913	0.973	0.978	0.985	0.980	0.975	0.985
0	0.899	0.967	0.974	0.982	0.976	0.969	0.981

Table 3. Accuracy of imputation for medium density chips for 4 classes of bulls with different numerator average relationships (AVgRel) with bulls in the reference population.

	High density chips of size greater 30k				
AVgRel(%)	8	11	15	108	208
>80	0.996	0.999	0.996	0.996	0.996
25 – 80	0.996	0.999	0.996	0.996	0.996
1 - 24	0.995	0.999	0.995	0.995	0.995
0	0.995	0.999	0.995	0.995	0.995
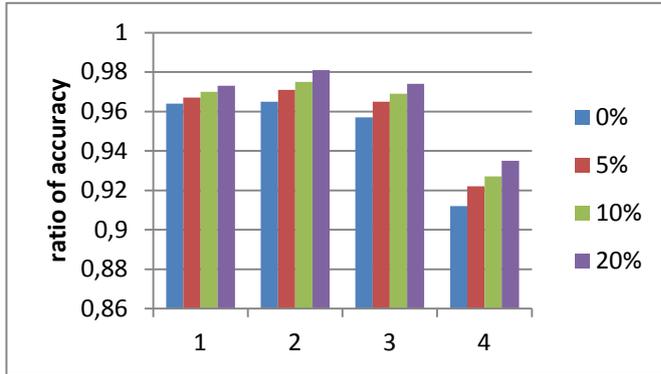
Figure 1. Ratio of milk yield genomic accuracy of imputed to original genotypes for chip 3 at 0, 5, 10 or 20% of polygenic effects for 4 classes of bulls defined in Table 2. Accuracies from original genotypes were 0.80, 0.84, 0.80 and 0.80 for the 4 classes of bulls.
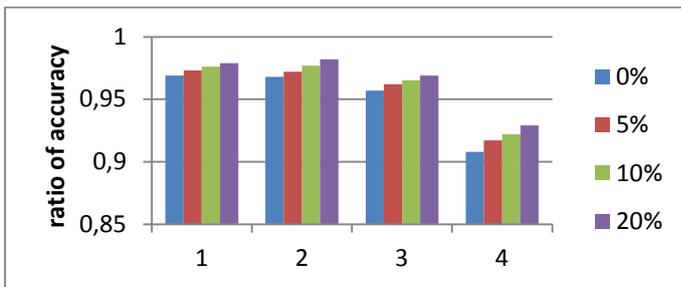
Figure 2. Ratio of milk yield genomic accuracy of imputed to original genotypes for chip 6 at 0, 5, 10 or 20% of polygenic effects for 4 classes of bulls defined in Table 2. Accuracies from original genotypes were 0.85, 0.84, 0.80 and 0.78 for the 4 classes of bulls .
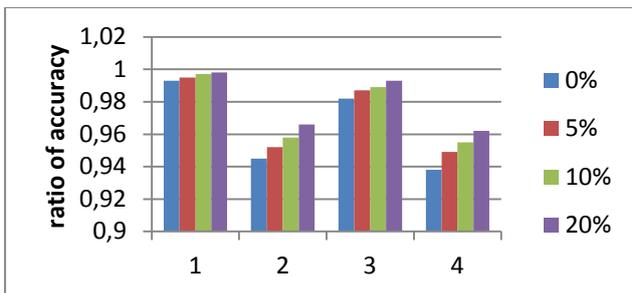
Figure 3. Ratio of milk yield genomic accuracy of imputed to original genotypes for chip 9 at 0, 5, 10 or 20% of polygenic effects for 4 classes of bulls defined in Table 2. Accuracies from original genotypes were 0.88, 0.81, 0.81 and 0.75 for the 4 classes of bulls .
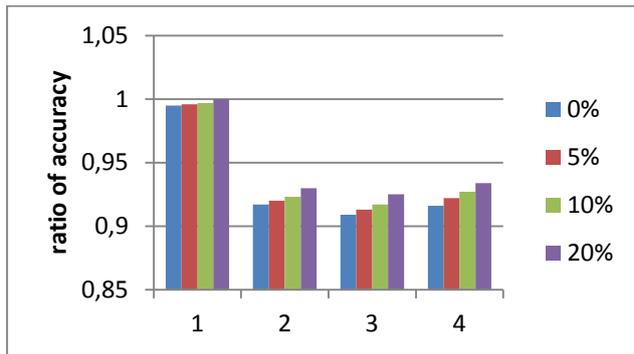
Figure 4. Ratio of SCC genomic accuracy of imputed to original genotypes for chip 3 at 0, 5, 10 or 20% of polygenic effects for 4 classes of bulls defined in Table 2. Accuracies of predictions from original genotypes were 0.73, 0.72, 0.70 and 0.67 for the 4 classes of bulls.
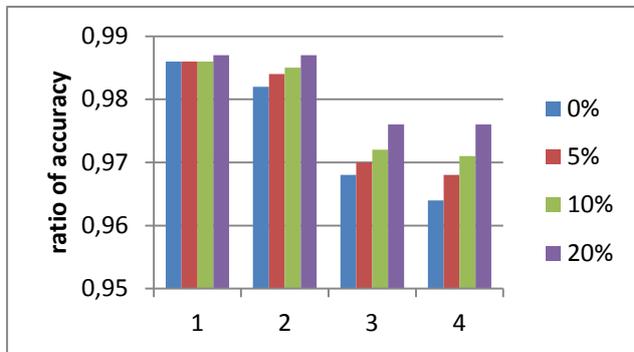
Figure 5. Ratio of SCC genomic accuracy of imputed to original genotypes for chip 6 at 0, 5, 10 or 20% of polygenic effects for 4 classes of bulls defined in Table 2. Accuracies from genotypes were 0.80, 0.73, 0.70 and 0.66 for the 4 classes of bulls.
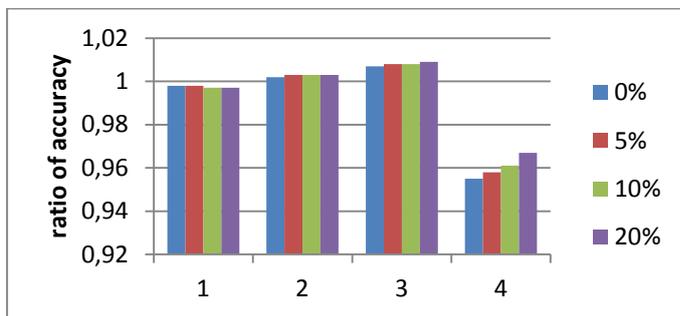
Figure 6. Ratio of SCC genomic accuracy of imputed to original genotypes for chip 9 at 0, 5, 10 or 20% of polygenic effects for 4 classes of bulls as defined in Table 2. Accuracies from original genotypes were 0.82, 0.75, 0.73 and 0.69 for the four classes of bulls.