Genomic Evaluation Using 50K and Imputed HD Genotypes in Guzera (*Bos indicus*) Breed

S. A. Boison[1], D. J. de A. Santos[2], J. F. Garcia[3], J. Sölkner[1], M. G. C. D. Peixoto[4] and M. V. G. B. da Silva[4]
[1]University of Natural Resources and Life Science, Vienna, Austria, [2]UNESP, Jaboticabal, São Paulo, Brazil, [3] UNESP, Araçatuba, São Paulo, Brazil, [4]Embrapa Dairy Cattle, Juiz de Fora, Brazil.

ABSTRACT: Lower and medium (Illumina 3K, 7K, 50K) density SNP markers have been shown to be less informative and explains a small proportion of the total additive genetic variance for most traits, in *Bos indicus* dairy and beef cattle breeds. The objective of this study was to estimate the proportion of additive genetic variance explained by actual Illumina 50K and imputed 777K (HD) genotypes in Guzera (*Bos indicus*) breed for milk (MY), fat (FY) and protein (PY) yield. We also studied the accuracy of genomic prediction. Imputation of 936 cows was done with FImpute using 75 sires. The proportion of the total additive genetic variance explained by markers increased from 0.62 (actual 50K SNPs) to 0.91 for the imputed HD. Imputed HD markers increased prediction accuracy by 13%, 2% and 10% for MY, FY and PY respectively, compared to the actual 50K.
Keywords: Additive variance; Accuracy; Imputation; FImpute; *Bos indicus*

Introduction

Practical application of genomic selection (GS) in the breeding industry depends among other factors, on the price of genotyping. The current medium or high density SNP chips are still expensive for widespread use considering the number of individuals needed to constitute sufficient reference population (RP) in order to achieve reasonable accuracies. To reduce the cost of GS for breeding programs, imputation approaches have been used (Khatkar et al. (2012); Badke et al. (2013); Hozé et al. (2013); VanRaden et al. (2013)). Imputation utilizes the genotyping of very few informative individuals on a medium to high density SNP panels and while the remaining individuals are genotyped on lower density SNP chips. Subsequently, imputation strategies are used to infer the un-typed markers. Lower (Illumina 3K, 7K), medium (Illumina 50K) or high (Illumina 777K) density SNP markers are particularly useful for genome wide association studies (GWAS) and GS. However, Illumina lower density or even medium density SNP markers have been shown to be less informative (ascertainment bias) and appear to explain a small proportion of the total genetic variance for most economically important traits in beef cattle in *Bos indicus* breeds (Elzo et al., 2013). In our current dataset of about 950 animals with Illumina 50K genotypes, initial analysis showed that, SNP markers only account for about 55-65% of the additive genetic variance for milk production traits. An obvious strategy to increase this proportion, and hence increase accuracies of GS and fine-tune regions of interest in GWAS studies is to use imputation. Thus, the objective of this study is to a) compare the estimate of the proportion of total additive genetic variance captured by actual Illumina 50K and imputed Illumina HD 777K genotypes in Guzera, a *Bos indicus* dual purpose cattle breed of Brazil; b) prediction accuracy of actual 50K and imputed HD genotypes; c) effect of imputation accuracy on the total additive genetic variance captured by SNP markers and estimated breeding value (EBV) rankings.

Materials and Methods

Phenotypic and Pedigree Data: Traits under study were deregressed estimated breeding values (dEBV, Garrick et al. (2009)), of 305 day Milk (MY, kg), Fat (FY, kg) and Protein (PY, kg) of both sires and dams. Variance components were estimated with the WOMBAT software (Meyer, 2007). The mean ± SD MY, FY and PY was 2142.92 ± 1032.82, 112.40 ± 57.38 and 70.50 ± 29.77 respectively. Pedigree data consisted of 6,039 animals (698 sires and 2558 dams). Heritabilities were estimated at 0.26, 0.28 and 0.26 for MY, FY and PY respectively. Reliabilities of 936 animals with genotypes and phenotypes ranged from 0.17 to 0.90 with an average of 0.47.

Genotypic data and Quality control: 75 sires were genotyped with the Illumina Bovine HD (777K) Beadchip and 973 cows were genotyped with the BovineSNP50 v2 BeadChip. Quality control (QC) was undertaken separately for the HD (777,962) and 50K (54,609). The following criteria were used for SNP quality control; SNPs mapped to the same position or with unknown positions were removed, SNPs with GenCall Score of <0.5 were set to missing, SNPs with minor allele frequency (MAF) < 0.02, call rate <95% and with exact p-value for HWE test $<10^{-6}$ were removed. Individuals with 10% missing genotypes were also removed from the data set. After QC, all the 75 animals remained with 508,334 for the HD whiles, 965 (8 deleted) cows with 28,546 SNPs remained. As stated above, only 936 animals with genotypes and phenotypes were used for subsequent analysis. The remaining SNPs from the QC of the 50K had 24,106 SNPs in common with the HD. Since cows were not genotyped on HD, imputation accuracy could not be assessed. Additionally, the effect of using the imputed HD genotypes for estimating marker effect, variance component and genomic EBVs (GEBV) could not be studied. Thus to try and mimic the potential effect of using imputed HD data on the estimates of variance component and GEBVs, the commercial Illumina 3K and 7K markers were subset from the actual 50K and imputed to 50K. We could thus directly estimate imputation accuracies using their actual 50K genotypes. The imputed 50K from the two scenarios (50K_3K

and 50K_7K) are then used for estimating variance component and genomic EBVs (GEBV). Rank correlations between the estimated breeding values from the actual 50K and imputed 50K are studied. Additionally, predictive ability of the imputed 50K are also reported (details below). The authors note here that, these 2 SNP chips have been shown to have lower imputation accuracy when imputed to 50K than imputation of 50K to HD (Berry et al. (2013)). Nonetheless, this strategy, gives us a fair approximation of the impact of the imputed data on our results. The Illumina 3K and 7K SNPs markers had 1,987 (91.8 % SNPs to be imputed) and 4,653 (80.7 % SNPs to be imputed) SNPs in common with the quality controlled 50K markers respectively.

Imputation: Genotype imputation was undertaken using the 75 sires with HD genotypes as RP. 965 cows were imputed from 50K (24,106 SNP markers) to HD (508,334). Additionally, as stated above, the 3K and 7K markers were imputed to 50K. FImpute v2 (Sargolzaei et al. (2012)) was used with prior pedigree information to link the RP and imputed set. Imputation accuracies were computed as the percentage of correctly called markers (%Ccall) and correlation (cor) between true 50K and imputed 50K_3K or 50K_7K markers. Genomic relationships (GRM) between the reference and imputed set were calculated using VanRaden (2008). Information from GRMs allows for easy comparison of imputation accuracies with other studies. Summary statistics on GRM were computed based on maximum (*relmax*), mean top 5 (*rel5*), 10 (*rel10*) and 20 (*rel20*) relationships for each individual.

Statistical analysis: Single trait genomic-polygenic model was fitted for MY, FY and PY to predict EBVs. The model was fitted on the actual 50K, imputed 50K_3K, imputed 50K_7K and imputed HD.

$$\mathbf{Y = Xb + Za + Wu + e}$$

Where Y is a vector of DrEBV for each traits on the genotyped animals; X is a vector of 1's, a is a vector of additive genetic effects with σ_a^2, Z is an incidence matrix coded as the dosage of the B allele ("2"). W is the incidence matrix and u is a vector of animal polygenic effects estimated from the pedigree data. Random polygenic effect u was assumed to follow $u \sim N(0, A\sigma_u^2)$; where A = Numerator relationship matrix. Random residuals e were assumed to be $\sim N(0, R\sigma_e^2); where\ R = {w_{ii}}^{-1};\ w_{ii} = r^2(1-r^2)^{-1}$.

The program GS3 (Legarra, 2009) was used to estimate variance components ($\sigma_a^2, \sigma_u^2, \sigma_e^2$) with the VCE option (MCMC; Number of iteration=250,000; Burn-in=50,000; Thinning = 50). The total genetic variance explained by all SNP markers (σ_g^2) was computed as $\sigma_a^2 \times \sum_{i=1}^{Nsnps} 2p_i(1-p_i)$, where p_i is the allele frequency of marker *i*. EBVs were computed for an individual *i* as $\hat{u}_i + z_i\hat{g}$. All animals were used for variance component estimation.

Predictability and Rank correlations: The dataset was split into training and validation set in an 8-fold cross validation procedure. The same model above was used to estimate marker effect from the training set and GEBVs in the validation set (MCMC sampling; Number of iteration=50,000; Burn-in=10,000; Thinning = 50). However, the newly generated variances from the earlier MCMC procedure were used as initial values. Predictability (accuracy) was calculated as correlation between dEBV and GEBV estimated with genomic polygenic model for imputed 50K (3K and 7K) and HD. Spearman rank correlation of GEBVs between the actual 50K data and imputed 50K were calculated as an indicator of the effect of imputation on GEBV rankings.

Results and Discussion

This study was aimed at estimating the proportion of total additive genetic variance explained by actual Illumina 50K and imputed 777K (HD) genotypes as well as its effect on genomic evaluations. Table 1 shows the estimated GRM within the reference and between the reference and imputed dataset. On average, the maximum relation for animals in the RP were around 0.42 (parent-offspring or full-sibs) while between reference and imputed set the relationship was lower (0.37). The relationship between imputation accuracy and GRM have been found to be positively correlated, although other factors like size of RP also plays a key role in imputation accuracy (Hozé et al. (2013)). We observed a strong linear relationship between ***rel5*** GRM and accuracy of imputation (results not shown).

Table 1: Average genomic relationships (GRM) within reference set (RP) and between RP and imputed set for 50K SNP chips

GRM	RP : RP	RP : Imputed
rel*max*	0.426 (0.117)	0.374 (0.135)
rel*5*	0.261 (0.090)	0.205 (0.070)
rel*10*	0.185 (0.071)	0.142 (0.051)
rel*20*	0.113 (0.049)	0.087 (0.034)

Imputation accuracies for the two tested scenarios of 50K_7K and 50K_3K are shown in Figure 1. Accuracies for imputing 3K to 50K were expectedly lower (87%; 0.89) than 7K to 50K (92%; 0.94). Weigel et al. (2010) reported similar imputation accuracy (90%) for Jersey bulls when they imputed approximately 3K SNPs to 48K. It is important to note that although 3K to 50K is similar to 50K to HD (92% vs. 95%) in terms of the number of un-typed markers imputed, results from VanRaden et al. (2013) and Berry et al. (2013) show much higher accuracies for 50K to HD than 3K to 50K or 7K to HD. Presumably, imputation accuracy for 50K to HD in this study is expected to be similar to or higher than the 7K to 50K. Imputation accuracies reported here, are much lower than in most other studies, this might be due to the small number of animals in the RP needed to build a comprehensive haplotype library as well as differences in population structure leading to different pattern of linkage disequilibrium. The effect of RP on impu-

Table 2: Posterior means and SD of genomic additive variance and proportion of total additive variance1 explained by markers for milk yield (MY, kg/105), Fat yield (FY, kg/103) and Protein yield (PY, kg/103).

Parameters	dataset	MY	FY	PY
σ_g^2	Actual 50K	4.418±0.402	1.686±0.122	1.394±0.095
	50K_3K	4.247±0.390	1.630±0.116	1.352±0.091
	50K_7K	4.382±0.396	1.664±0.117	1.373±0.093
	Imputed HD	16.190±0.802	13.750±0.682	13.364±0.703
$\sigma_g^2/(\sigma_g^2+\sigma_u^2)^1$	Actual 50K	0.670±0.047	0.590±0.025	0.592±0.022
	50K_3K	0.666±0.045	0.582±0.025	0.584±0.022
	50K_7K	0.681±0.046	0.588±0.024	0.588±0.022
	Imputed HD	0.960±0.013	0.879±0.011	0.880±0.011

σ_g^2 – total additive genetic variance captured by all markers; σ_u^2 – polygenic variance estimated with pedigree information.

tation accuracies have been comprehensively discussed by Badke et al. (2013) and Hozé et al. (2013).

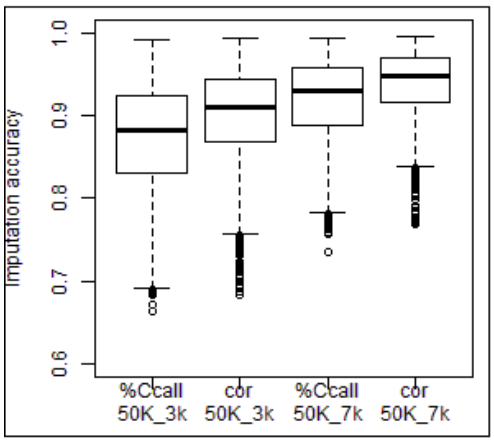

Figure 1: Boxplot of imputation accuracy from 3K and 7K to 50K

Proportion of total additive variance captured by SNP markers increased from 0.67 to 0.97; 0.59 to 0.88 and 0.59 to 0.88 for MY, FY and PY respectively (Table 2). Although imputation accuracies were lower than 93%, this seems to only affect the estimation of variance components slightly as shown in Table 2 for imputed 50K with either 3K or 7K. Similar trend was also observed by Weigel et al. (2010) in scenarios where imputation accuracies were about 90%. Imputation accuracies <90% resulted in a huge under-estimation of variance component. Results from the spearman rank correlation between GEBV estimated from imputed 50K (from 3K or 7K) and actual 50K are high (>0.94). These results are important because it validates extension to imputed HD genotypes which can be used for GEBV estimation with little effect on ranking of animals. Elzo et al. (2013) also reported similar spearman rank correlations of >93% for genomic-polygenic models studying Brangus heifers using actual and imputed 50K.

Genomic predictions using imputed HD genotypes were slightly more accurate (8% averaged across all traits) than using actual 50K genotypes (Table 4). Weigel et. al (2010) and Vazquez et al. (2010) all reported higher accuracies of prediction using denser SNP markers. There was a small loss in accuracy of prediction using imputed 50K from 3K, while prediction accuracy remained the same for the imputed 50K from 7K scenario (Table 4). The higher accuracies observed for all the traits with imputed HD, might be due to the resultant increase in marker variance and precise estimation of marker effect using the imputed HD. Accuracy was similar to the results of Erbe et al. (2012) on Jersey sires using a GBLUP model for MY, FY and PY. Prediction accuracy increases for

Table 3: Spearman rank correlation between dEBVs estimated with actual 50K SNPs and imputed 50K (from 3K and 7K) and Imputed HD for milk yield (MY, kg), Fat yield (FY, kg) and Protein yield (PY, kg)

Scenario	Correlation	MY	FY	PY
50K_3K	GEBV; 50K	0.952	0.945	0.946
	G-PEBV; 50K	1.000	0.998	0.998
50K_7K	GEBV; 50K	0.961	0.956	0.957
	G-PEBV; 50K	1.000	1.000	1.000
Imputed HD	GEBV; 50K	0.994	0.986	0.986
	G-PEBV; 50K	1.000	0.994	0.993

G-PEBV = GEBV-Genomic breeding values estimated with SNP markers; PEBV-Pedigree breeding values; p-value were highly significant (depending on scenario $>10^{-3}$).

Table 4: Accuracies of genomic predictions for milk yield (MY, kg), Fat yield (FY, kg) and Protein yield (PY, kg) using actual 50K, imputed 50k (from 3k and 7k) and imputed HD

dataset	MY	FY	PY
Actual 50K	0.39 ± 0.10	0.43 ± 0.08	0.41 ± 0.08
50K_3K	0.36 ± 0.09	0.41 ± 0.08	0.40 ± 0.07
50K_7K	0.41 ± 0.10	0.43 ± 0.08	0.41 ± 0.07
Imputed HD	0.44 ± 0.03	0.44 ± 0.02	0.45 ± 0.03

HD genotype data compared to 50K mostly when there is a corresponding increase in phenotypes. Thus, genomic prediction models that combine genotypes with pedigree like the single-step GBLUP can be used to maximize gain in accuracy. K-fold cross validation was used because, correlation of inaccurate EBVs (below 0.2 in this study) to

GEBVs might lead to incorrect prediction accuracy. A K-fold cross validation allows each individual to be predicted once. We do note that, there can be slight overestimation of genomic prediction accuracy with this kind of validation.

Conclusion

The proportion of total additive genetic variance captured by SNP markers increased by about 49% (average across all traits; MY, FY and PY) using imputed HD genotypes. Using imputed genotypes in genomic prediction models, the estimation of GEBV and ranking of animals was only slightly affected by imputation accuracy. Imputed HD genotypes increased prediction accuracy by about 8% on average.

Literature cited

Badke, Y.M., Bates, R.O., Ernst, C.W., et al. (2013). BMC Genetics. 14, 8.

Berry, D.P., Mullen, M.P., and Cromie, A.R. (2013). Proc. Assoc. Advmt. Anim. Breed. Genet 20, 542–545.

Garrick, J. D., Taylor, J. F., Fernando, R. L. (2009). Genetics Selection Evol. 41, 1-8.

Elzo, M. A., Thomas, M.G., Martinez, C. A., et al. (2013). Livest. Sci. 159, 1–10.

Erbe, M., Hayes, B.J., Matukumalli, L.K., et al. (2012). Journal of Dairy Sci. 95, 4114–4129.

Hozé, C., Fouilloux, M.-N., Venot, E., et al. (2013). Genetics Selection Evol. 45, 33.

Khatkar, M.S., Moser, G., Hayes, B.J., et al. (2012). BMC Genomics 13, 538.

Legarra, A. (2009). GS3 ⟨ http://snp.toulouse.inra.fr/ ~alegarra/manualgs3_2.pdf⟩ .

Meyer, K. (2007). WOMBAT: J. Zhejiang Univ. Sci. B 8, 815–821.

Sargolzaei, M., Chesnais, J.P., and Schenkel, F. (2012). Open Ind. Sess. Oct. 30, 2012, 1–10.

VanRaden, P.M. (2008). J. Dairy Sci. 91, 4414–4423.

VanRaden, P.M., Null, D.J., Sargolzaei, M., et al. (2013). Journal of Dairy Sci. 96, 668–678.

Vazquez, A.I., Rosa, G.J. M., and Weigel, K.A., et al. (2010). Journal of Dairy Sci. 93, 5942–5949.

Weigel, K. A., de los Campos, G., and Vazquez A. I., et al. (2010). Journal of Dairy Sci. 93, 5423–5435.