ASSESSING IMPUTATION ACCURACY USING A 15K LOW DENSITY PANEL IN A MULTI-BREED NEW ZEALAND SHEEP POPULATION

R.V. Ventura[1,2], M. Lee[3], S.P. Miller[1,4], S.M. Clarke[4]and J.C. McEwan[4]

[1]Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, Ontario, Canada
[2] Beef Improvement Opportunities, Guelph, Ontario, Canada
[3] Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand
[4]Invermay Agricultural Centre, AgResearch Limited, Mosgiel, New Zealand

SUMMARY

Imputation has enabled genomic selection in commercial livestock, taking advantage of a more cost effective Low Density (LD) panel, increasing the number of genotyped animals and hence accelerating the adoption process. A 5K LD panel has been employed commercially in New Zealand. This study investigated the accuracy of imputation to 50K and High Density (HD) panels using a new 15K panelbeing developed by the International Sheep Genomics Consortium in four scenarios across two multi-breed New Zealand sheep populations. Theprototype panel resulted in higher values of imputation accuracy compared with the current LD panel (5K), which willbenefit the implementation of genomic selection for the sheep industry in New Zealand.

INTRODUCTION

Imputation is a robust tool able to infer the genotype at anun-genotyped locus and has been largely adopted for minimizing costs of genotyping in livestock breeding including sheep in New Zealand. Imputation assessment using the 5K LD panel in sheep was previously reported by Australian researchers (Hayes *et al.* 2012). In the New Zealand sheep industry, application of the current version of the low density panel (5K) has identified several genomic regions where imputation accuracy could be improved (Ventura et al. 2015, paper in submission), which could increase the accuracy of genomic predictions and further improve the identification of regions associated with traits of economic importance.A new 15K panel (in the process of design),containing markers selected by the International Sheep Genome Consortium (ISGC - www.sheephapmap.org),was used in this study to investigate imputation accuracies from 15K to both 50K and High Density (HD) panels in a multi-breed sheep population, pointing to potential regions for improvement over the 5K LD panel,which is used commercially for genomic selection in New Zealand sheep.

MATERIALS AND METHODS

Population imputation was implemented using the FIMPUTE 2.2 software(Sargolzaei *et al.* 2014). A total of 15,443 animals,part of the Beef and Lamb NZ genetics program (formally Ovita), composing a multi-breed sheep population, were genotyped with the Illumina OvineSNP50 Genotyping BeadChip(53,903 markers) and used in the present study to investigate the imputation from the low density panels (LD) 5K and 15K to the 50K panel. A second group of animals, part of the FarmIQ project, were genotyped using the Ovine Infinium® HD SNP BeadChip (606,006 markers) and were used to carry out the imputation from 15K to the HD panel. The HD animals were selected from eight flocks predominately of terminal composite breeds. Many of the animals were from recent breed developments (http://www.focusgenetics.com/) with undefined breed ratios and are best described as composites. The majority of the animals (~97%) in the second group were born in the period 2010 to 2013. The total number of animals used in the terminal composite population was 2,868, where 300 of these were used as a validation set and were born in 2013.The same strategy of using the youngest animals to be imputed was applied for the

population with 50K genotypes. Only autosomal markers were included in this investigation.For the imputation from LD to the 50K level, 12,853 markers (referred to 12K subsequently) out of the new 15K LD panel, remained after quality control.For the imputation from 15K to HD, 14,844 markers from the new LD set of SNPs were located on the HD panel and remained after quality control. Table 1 shows seven scenarios: six covering the imputation from 5K and 12K to the 50K panel in Romney, Coopworth and Perendale animals (Scenarios 1_R, 2_C and 3_P, respectively) and an additional scenario (4_TC) investigating the imputation from 15K to HD in the terminal composites. All LD panels used in this study were simulated by keeping markers in common between the respective LD and higher density panels and deleting remaining markers exclusively located in the higher density set (50K or HD).

Table 1. Description of imputation scenarios from 5K and 12K to 50K, and from 15K to HD panel

Scenarios	No. reference animals	No. imputed animals	Reference animals description	Imputed group breed[2]	Density
1_R	4256	1000	Romney	Romney	5K &12K to 50K
2_C	15443	250	All breeds	Coopworth	5K & 12K to 50K
3_P	15443	250	All breeds	Perendale	5K & 12K to 50K
4_TC	2568	300	Terminal composite breed	Terminal composite breed	15K to HD

Imputation accuracy was investigated using the allelic squared Pearson correlation (r^2) and concordance rate (CR), determined as the proportion of the correctly imputed markers out of all markers that were inferred after imputation. In both cases, the imputed and true genotypes (before deletion to build the LD panel) were compared. Common SNPs between LD and HD panels (15K) were not considered during the imputation accuracy determination.

RESULTS AND DISCUSSION

The accuracy of imputation from 5K to 50K ranged from 87.89% to 89.97% using the concordance rate measure and from 65.42% to 68.22% when the r^2 per SNP was calculated (Table 2). Concordance rate, calculated per animal or SNP, provides the same accuracy. Accuracies determined using r^2 per SNP marker,as done in this study,are usually lower than values calculated based on the animal, mainly due to the number of markers that are taken in consideration for the correlation estimates. An average gain in accuracy of 5.68% and 16.07% in CR and r^2, respectively, was noted after using the new 12K panel as the LD panel rather than the current 5K. The imputation from 15K to HD, performed in the second group of animals (terminal composite group), resulted in a CR imputation accuracy of almost 98% and squared correlation (r^2)of 88.70%.

Table2: Imputation accuracy under different scenarios

Scenario	5KCR	5Kr^2	12KCR	12Kr^2	15KHDCR	15KHDr^2
1_R	89.07	67.06	94.77	83.24	-	-
2_C	89.94	68.22	94.92	82.96	-	-
3_P	87.89	65.42	94.26	82.70	-	-
4_TC	89.03	-	-	-	97.81	88.70

The accuracy of imputation (CR) per animal, from 5K and 12K to 50K, is presented in Figure 1-left. Accuracies for almost all individuals were substantially increased by adding markers in the sparser panel and the largest gains in accuracy using the 12K panel (5K + new markers) were obtained for animals that obtained the lowest accuracies with the sparser panel (5K). Figure 1-right shows accuracy of imputation from 15K to HD where all individuals had their missing genotypes inferred with at least 90% success.

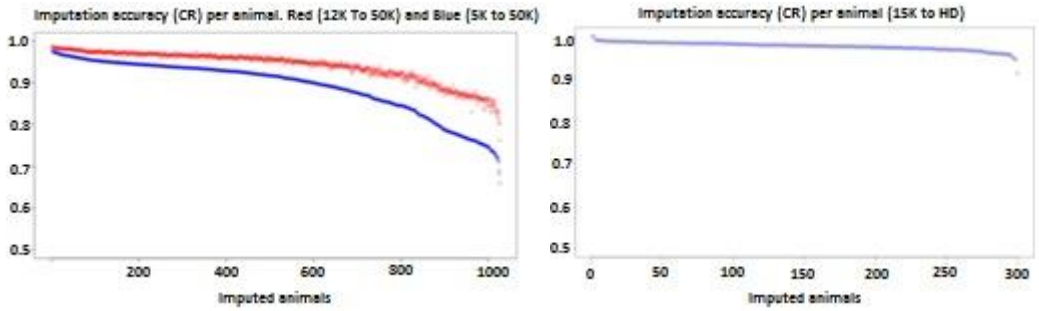

Figure 1. Accuracy of imputation per animal: Left – imputation from 5K(blue) and 12K(red) to 50K panel in Romney animals (Scenario 1_R). Right: imputation from 15K to HD panel in terminal composite breed. (In both plots X is reported as number of imputed animals and Y, as the CR measure of imputation accuracy ranging from 0.5 to 1.0).

A considerable increase in imputation accuracy to 50K for the 12K panel compared to the previous 5K panel was observed as illustrated in Figure 2. Almost all regions had higher imputation accuracy imputing from the 12K panel as compared to the 5K considering both r^2 and CR as metrics. The first 20Mb illustrates a region where the accuracy is improved considerably with the 12K panel. As illustrated in Figure 2, CR imputation accuracy per marker was higher than r^2 for all three scenarios. As reported by several authors in other species(Bouwman and Veerkamp 2014; Sargolzaei *et al*. 2014),imputation of markers at low MAF have lower r^2 accuracy than regions with higher MAF as can be noted by comparison of r^2 accuracy associated with MAF across different regions. The same pattern of increased imputation accuracy can be noted in the last two plots in Figure 2, where the increased accuracy showed almost the same trend for the same regions even in different populations and with higher density panels.

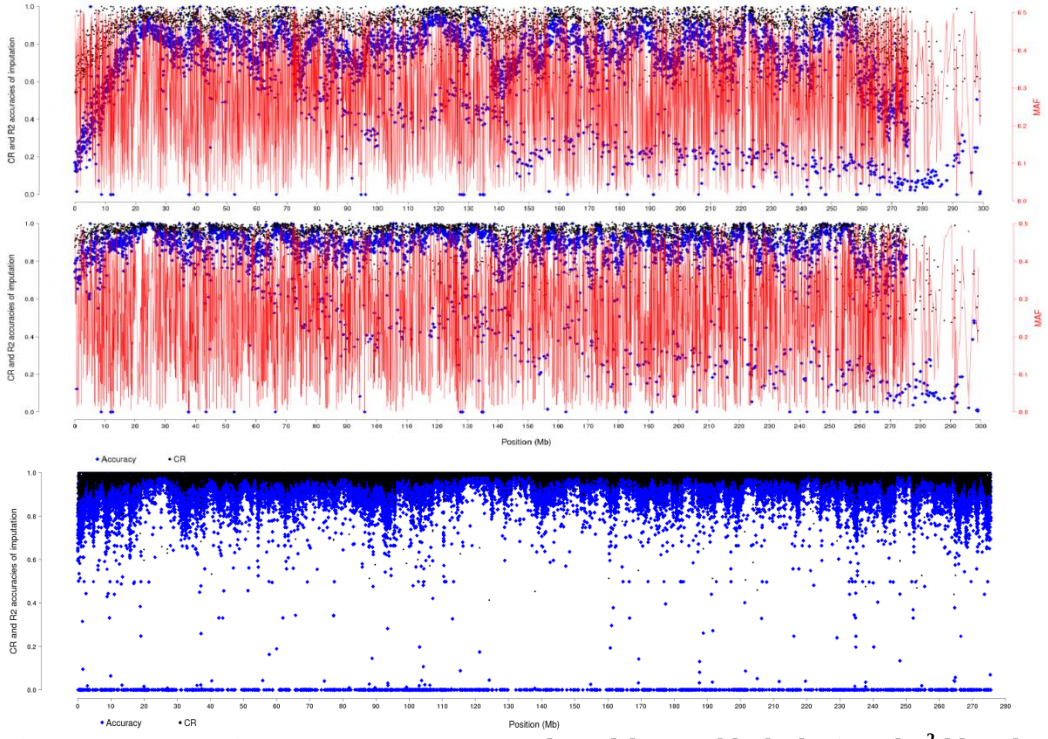

Figure 2 – Imputation accuracy per SNP evaluated by CR(black dots) and r^2(blue dots) according to the minor allele frequency (MAF is represented by the red line). First plot on top shows imputation accuracy from 5K to 50K for Chr1 and the central plot investigated imputation from 12K to 50K in Romney animals (Scenario 1_R). Last plot on the bottom shows imputation from 15K to HD in a terminal composite breed (Scenario 4_TC). (X is reported as Position (Mb) and Y, as the CR and r^2 measures of imputation accuracy ranging from 0.5 to 1.0)

The new 15K panel is still under development by the ISGC and the test results of this prototype panel will be used to inform the final panel implemented. Better imputation accuracy, especially at the chromosome ends can be expected with the new panel when implemented in New Zealand sheep.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the following organizations: Beef and Lamb New Zealand Genetics for access to 50K genotypes and FarmIQ (Ministry for Primary Industries' Primary Growth Partnership fund), for access to the HD genotypes, New Zealand sheep breeders for submitting samples, AgResearch Animal Genomics staff for genotyping and the ISGC for providing the prototype 15K panel design.

REFERENCES

Bouwman A. C. and Veerkamp R.F. (2014) *BMC Genet.* **15**:105.

Hayes B. J., BowmanP. J.,DaetwylerH. D.,Kijas J. W., and van der WerfJ.H.J.(2012) *Anim. Genet*. **43**:72.

Sargolzaei M., Chesnais J.P. and Schenkel F.S.(2014) *BMC Genomics* **15**:478.