Proc. Assoc. Advmt. Anim. Breed. Genet: 24: 267-271

EXPLORING IMPUTATION ACCURACY ACROSS THE BOVINE X CHROMOSOME

T.V. Nguyen[1], S. Bolormaa[1], C.M. Reich[1], A.J. Chamberlain[1], A. Medley[2], C. Schrooten[3], H.D. Daetwyler[1,4] and I.M. MacLeod[1]

[1] Agriculture Victoria, AgriBio, Centre for AgriBiosciences, Bundoora, Victoria, 3083, Australia
[2] CRV, Hamilton, New Zealand
[3] CRV, Arnhem, Netherlands
[4] School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, 3083, Australia

SUMMARY

Many of the current imputation benchmarking studies are performed on autosomes with limited studies addressing the X chromosome. Furthermore, the X chromosome genome map has recently been updated in the new ARS-UCD1.2 bovine reference genome. In this study, we evaluated the empirical accuracy of imputation from a low-density SNP array (LD) to 50K and then high-density (HD) for the pseudo-autosomal region (PAR), non-PAR, and autosomes across several scenarios using multiple dairy breed groups. Overall, imputation accuracies for the PAR were very low when imputing from LD to 50K, while accuracy for the non-PAR was comparable with autosomes. We demonstrated that imputation accuracies for the PAR increased when the PAR & non-PAR were merged for imputation. However, while this strategy performed well for imputing LD to 50K, there was no advantage when imputing from real 50K genotypes to HD. In addition, when imputing all chromosomes to HD level, imputing from real 50K to HD resulted in an overall higher accuracy than imputing from LD to 50K to HD, with the PAR region showing the most improvement. By separately imputing only the end segment of five autosomes and comparing accuracy with the PAR region, we demonstrated that the PAR region is more difficult to impute accurately, perhaps due to higher recombination rates. Therefore, future SNP genotyping panels should have SNP density in the PAR at least equivalent to that of the 50K SNP panel to achieve a good imputation result.

INTRODUCTION

Genomic selection (Meuwissen *et al*. 2001) has created a dramatic breakthrough in the dairy industry during the last two decades. Accurate prediction of breeding values requires medium to high-density genome-wide markers, but many of the dairy genomic reference populations have been genotyped on a range of lower density platforms (6,000 to 25,000 markers) to reduce costs. Genotype imputation is considered an effective approach to provide the marker density required by the industry. To date, most studies that examined the empirical accuracy of imputation from low-density (LD) to medium (50K) or high-density (HD) SNP genotypes investigated imputation of autosomes only and this is generally highly accurate (Calus *et al*. 2014). The X chromosome generally requires modifications to the imputation pipeline because it has a 5.7 Mb region of homology between chromosome X & Y called the pseudo-autosomal region (PAR) and a larger non-PAR that is haploid in males. Two studies investigated the accuracy of imputation on the X chromosome (LD to 50K) and found it was much less accurately imputed compared to autosomes in cattle (Su *et al*. 2014; Mao *et al*. 2016). However, the imputation of the X chromosome warrants further study for three key reasons. First, these studies used the UMD-3.1 reference genome map, while recently the X chromosome map has been extensively updated on the ARS-UCD1.2 bovine reference, in particular the PAR region (Figure 1A). Second, these authors tested imputation to 50K density only and did not investigate strategies to improve the PAR imputation accuracy. Third, there may be important genetic variation on the X chromosome for economically important traits as reported for fertility (Pacheco *et al*. 2020). In this study, we evaluated the empirical

accuracy of imputation from a LD SNP array to 50K and then to HD for the pseudo-autosomal region (PAR), non-PAR and autosomes across several scenarios using multiple dairy breed groups.

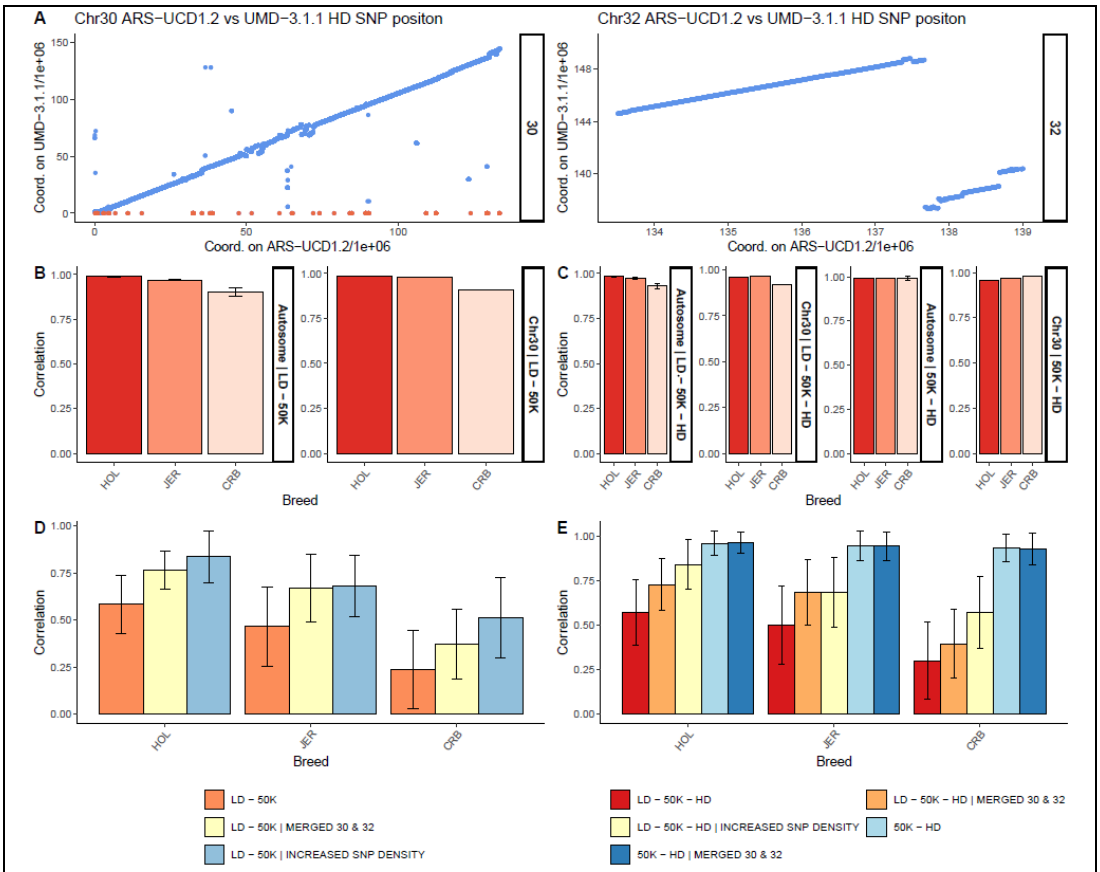

Figure 1. (A) Bovine HD SNP coordinates on Chr30 (Left) and Chr32 (Right) on ARS-UCD1.2 and UMD-3.1.1 reference genome. Red dots represent SNPs that were relocated from other chromosomes. (B) Average imputation accuracy for autosomes and Chr30 imputed from LD to 50K, error bars represent Standard deviation (SD) of accuracies across autosomes. (C) Average imputation accuracy for autosomes and Chr30 imputed to HD level from either LD (via 50K) or direct from real 50K. Error bars represent SD of accuracies across autosomes. (D) Imputation scenario accuracy for Chr32 at 50K level. Error bars represent SD of accuracy across SNP. (E) Imputation scenario accuracy for Chr32 imputed to HD either from LD (via 50K) or direct from real 50K. Error bars represent SD of accuracy across SNP.

MATERIALS AND METHODS

The target animals used for this study included 35 Jersey (JER), 35 Holstein (HOL), and 35 crossbred (HOL, JER) bulls (CRB) and were genotyped using the Illumina® BovineHD chip. GenCall threshold score was set at 0.6: animals and SNP were removed if >10% of genotypes fell below this threshold. The marker map positions were based on the ARS-UCD1.2 reference genome (Rosen *et al*. 2020). The boundary between the non-PAR and PAR (hereby noted as Chr30 and Chr32 respectively) was set to 133,300,518bp (Johnson *et al*. 2019). Chr30 and Chr32

were imputed separately unless otherwise stated. We masked the HD genotypes (714,452 SNPs) to simulate either a LD SNP-chip of 7,135 markers or the 50K chip (40,397 markers). Two sub-experiments were conducted: (1) All autosomal (Chr1 to 29), non-PAR (Chr30), and PAR (Chr32) LD genotypes were either imputed to 50K and then to HD level or from real 50K genotypes to HD. (2) For comparison between accuracy of imputation on the PAR and autosomes, we selected the last 5,708,563 bp segment (equivalent to the length of Chr32) on Chr 1,2,3,4 and 5 and re-imputed only these short segments. On these autosomal segments the LD SNP density (N≈30) was double that of the PAR, therefore we compared imputation at two SNP densities: first we reduced every other marker of the autosome sets to mimic the density on Chr32 (N=15), and second, we doubled the density on Chr32 by including several 50K variants to mimic the LD marker density on autosomal segments (N=30). Imputation was performed using FImpute V3.0 (Sargolzaei *et al*. 2014). We estimated the accuracy of imputation as Pearson's correlation coefficient (r) between imputed and real genotypes and results are reported based on the mean per-SNP accuracy. Imputation to 50K was performed with a reference set of 14,000 animals that included HOL and JER, and imputation to HD was conducted with a similar mixed breed reference of 2,700 animals.

RESULTS AND DISCUSSION

In this study, we tested several imputation strategies for the PAR & non-PAR on the X chromosome and compared the accuracy to that of the autosomes. At 50K level, we found that pooling all samples (HOL, JER, and CRB) and using a mix breed reference gave similar imputation accuracy compared to imputing HOL or JER target sets separately with only the same breed in the reference. Therefore, we present results using pooled target and reference sets but show average accuracies for each breed group. We found some differences in accuracies between the breed groups: the CRB were lowest for LD to 50K (Fig. 1B) but as high as HOL and JER when imputing from real 50K to HD (Fig. 1C). However, the CRB were more related to the smaller HD reference than the 50K reference, implying that this caused the variation in imputation accuracies, to confirm this, we masked the HD reference down to 50K level to act as a new 50K reference and found similar imputation accuracy for all three breed groups (~0.96) .

We found that Chr30 imputation accuracy was high (>0.97) and comparable to autosomes for both 50K (Figure 1B) and HD level across target breed sets (Figure 1C), indicating that it is useful to include imputed genotypes from the non-PAR for downstream analysis. Conversely, Chr32 imputation accuracy was very low when imputing from LD to 50K (Figure 1D). Although it is recommended that Chr30 and Chr32 are imputed separately, by combining Chr32 and Chr30 (and re-extracting Chr32 genotypes) the accuracy increased by at least 15% for all breed groups when imputing to 50K (Figure 1D). Nonetheless, the accuracy is still rather low for downstream analysis. It should be noted that this strategy slightly reduced the imputation accuracy on Chr30 (results not shown), so markers on Chr30 should be imputed separately. Per SNP statistics for Chr32 showed that accuracy was improved in the borderline region between Chr30 & Chr32. This strategy of merging Chr30 and Chr32 provides a practical approach for historical datasets with low-density genotypes because increasing SNP density is not an option but should also be tested in females because our target animals were all males. When the SNP density was doubled on Chr32 (15 to 30) to mimic the number of SNPs in the last 5.7 Mb segment of Chr 1,2,3,4 and 5 the accuracy increased further (Figure 1D). At HD level, imputation of Chr32 from real 50K genotypes was always more accurate (0.92-0.97) than imputation from LD regardless of scenario (0.29-0.84). Although this was a little lower than the accuracy for Chr30 and autosomes, it was of high enough quality for downstream analyses. Notably, there was no longer an advantage in merging Chr32 with Chr30 for imputation from real 50K to HD (Figure 1E). This contradicts the result observed for LD to 50K level, suggesting that the denser markers available on the 50K SNP chip on Chr32 (99 SNPs) enable good resolution of Chr32 haplotypes.

One potential reason for the low accuracy on Chr32 may be simply that it is a very short segment to impute, and typically on all chromosomes the imputation accuracy tends to drop at the ends of the chromosomes. However, the results of our autosomal segment imputation test, demonstrated that when the marker density was made equivalent (either reducing density on the autosomal segments 1-5 or increasing density on Chr32), the accuracy was always better for the autosomal segments relative to Chr32 (Figure 2). We believe that higher recombination frequencies on Chr32 compared to the autosomes (Van Laere et al. 2008) might be responsible for increased haplotype complexity of this region. Therefore, when designing SNP panels for genotyping, it is perhaps critical to use SNP densities on Chr32 that are at least equivalent to those on the 50K chip.

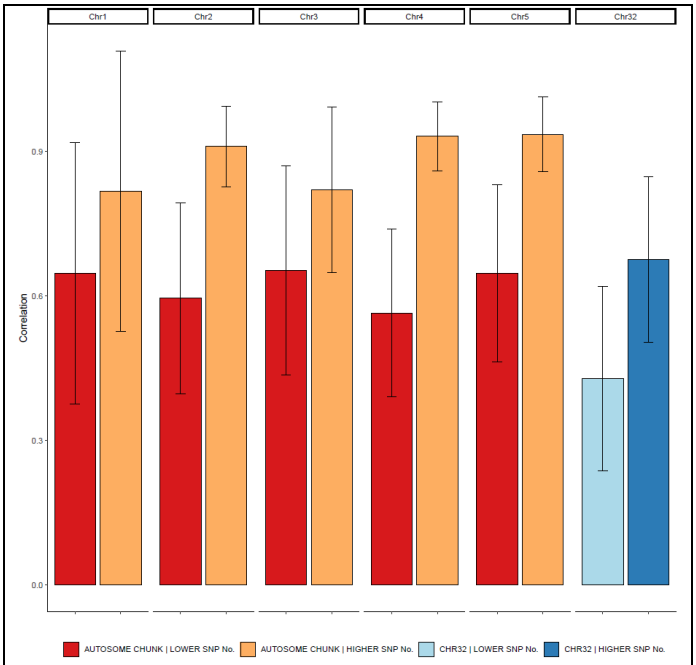

Figure 2. Imputation accuracy for markers at two LD densities imputed to 50K on a 5.7 Mb end segment of Chromosome 1,2,3,4 and 5, compared with markers on the PAR region (Chr32) Results were pooled across three breed groups & error bars show SD across SNPs.

CONCLUSIONS

This study compares accuracy of imputation for autosomes and the X chromosome including several imputation scenarios for the PAR on bovine genome ARS-UCD1.2. We demonstrated that the accuracy of PAR imputation can be improved from LD to 50K by imputing the PAR & non-PAR together and re-extracting the PAR markers. However, if designing new SNP genotyping panels, we recommend SNP density in the PAR should be equivalent to that of the 50K SNP panel because this can greatly increase imputation accuracy.

ACKNOWLEDGEMENTS

DairyBio, a joint venture project between Agriculture Victoria (Melbourne, Australia), Dairy Australia (Melbourne, Australia) and the Gardiner Foundation (Melbourne, Australia), funded resources used in the analysis.

REFERENCES

Calus MPL, Bouwman AC, Hickey JM, Veerkamp RF, Mulder HA. (2014) *Animal* **8**: 1743.
Johnson T, Keehan M, Harland C, Lopdell T, Spelman RJ, Davis SR, Rosen BD, Smith TPL, Couldrey C. (2019) *J. Dairy Sci.* **102**: 3254.
Mao X, Johansson AM, Sahana G, Guldbrandtsen B, De Koning D-J. 2016. *Genetics* **157**: 1819.
Pacheco HA, Rezende FM, Peñagaricano F. (2020). *J. Dairy Sci.* **103**: 3304.
Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin

A, Couldrey C et al. (2020) GigaScience 9, doi: 10.1093/gigascience/giaa021
Sargolzaei M, Chesnais JP, Schenkel FS. (2014). *BMC Genomics* **15**: 478.
Su G, Guldbrandtsen B, Aamand GP, Strandén I, Lund MS. (2014) *Genet. Sel. Evol.* **46**: 47.
Van Laere A-S, Coppieters W, Georges M. (2008) *Genome Res.* **18**: 1884.