CONSIDERATIONS IN IMPUTATION OF SKIM SEQUENCED DATA

M.S. Tahir[1], J. Wang[1], A.J. Chamberlain[1,2], C.M. Reich[1], B.A. Mason[1] and I.M. MacLeod[1,2]

[1] Agriculture Victoria, Centre for AgriBioscience, Bundoora, VIC, 3083 Australia
[2] School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3083 Australia

SUMMARY

Skim sequencing genotypes have become a cost-effective alternative to standard SNP array genotypes for genomic studies. However, compared to deep sequencing, skim sequencing suffers from randomly missing or inaccurate genotypes. The true genotypes can often be recovered with imputation using high-quality sequenced reference population genotypes. We skim-sequenced 293 cattle at 1x depth at an estimated cost of 36 AUD per sample. We aligned the skim sequence data of each animal to the bovine reference genome and imputed it to the whole genome sequence using GLIPMPSE2. Two imputed genotypes metrics (dosage: probability of called genotype and best-guess: 0, 1, or 2), called by GLIMPSE2, were compared to the real genotypes of the same animals with custom XT-50K SNP bead-chip to calculate per-SNP and per-animal imputation accuracy. The per-SNP and per-animal imputation accuracy with dosage genotypes were found to be 0.96 and 0.97, respectively. Sequencing depth of whole genome, imputation chunks, and 5kb windows around variants was determined for all animals and compared to the per-animal imputation accuracy. The correlation between the mean genome sequencing depth and imputation accuracy (per-animal) across all animals was weak (0.04). However, the difference in the mean sequencing depths of the whole genome, imputation chunks, and 5kb windows around variants between the animals with highest and lowest imputation accuracy was significant. The findings of this study show that skim sequence genotypes imputed with deeper sequenced reference population were highly accurate and cost-effective. Optimising the depth of skim sequencing can further improve the imputation accuracy.

INTRODUCTION

Skim sequencing, also known as low-coverage (< 3X) whole genome sequencing, has emerged as a cost-competitive alternative to standard SNP array genotyping (Li *et al.* 2021). It can capture the existing array-based SNPs as well as new variants in individual animals. A disadvantage of skim sequencing is missing information for some random genomic regions. This results in either sporadic missing genotypes or inaccurate genotypes. Missing information in skim sequence can be recovered with imputation using a high-quality deep-sequenced reference population (Pasaniuc *et al.* 2012).

Imputation methodology and landscape of sequencing depth can impact the final accuracy of the skim-sequence imputed genotypes (Al Bkhetan *et al.* 2019). For example, Beagle (Browning and Browning 2007) is highly accurate for imputation of variant genotypes to whole-genome sequence but showed relatively low imputation accuracy when used with skim-sequenced samples (Daetwyler *et al.* 2021) since it was not developed for such data. New imputation tools like GLIMPSE2 (Rubinacci *et al.* 2023) and QUILT (Davies *et al.* 2021) have been developed specifically to impute skim-sequenced data. Both tools are reported to show similar imputation accuracy, but GLIMPSE2 is computationally more efficient compared to QUILT (Rubinacci *et al.* 2023).

In livestock, it is important to maintain continuity between new and old genotyping platforms so that genetic evaluations relying on a particular variant panel are not negatively affected. In this study, GLIMPSE2 was used to impute skim-sequenced data of 293 cattle to the whole genome sequence (WGS) level. The empirical accuracy of imputation was tested for the variants that overlapped a custom XT-50K variant panel used for genetic evaluations in the Australian dairy industry (Xiang *et* *al.* 2021). We also tested the impact of skim-sequencing depth of target samples on imputation accuracy.

MATERIALS AND METHODS

A total of 293 (142 Holstein, 148 Jersey, and 3 crossbred) animals were skim-sequenced to an average 1x coverage costing 36 AUD per sample. Sequences were aligned to the bovine reference genome using BWA (Li *et al.* 2009). The aligned bam files were used as input for the imputation pipeline. Run8 of the 1,000 bull genomes project was used as the imputation reference with 4,118 animals of various *Bos. taurus* breeds. GLIMPSE2, with default parameters, was used to impute genotypes for 39,457,248 variants. It splits the reference genotypes into imputation chunks, phases the target samples using pre-phased information from the split chunks of the reference panel, imputes the missing genotypes, and then ligates the imputed chunks of the target samples. We also tested non-default values for parameters like imputation chunk size 18 Mb (default=4 Mb), and effective population sizes 1,000 and 10,000 (default=100,000).

GLIMPSE2 provides two metrics of imputed genotypes: 1) dosage, the estimated alternative allele dosage (a continuous metric of probability of genotype called), and 2) best-guess, the genotype called (0, 1, or 2) based on estimated allele dosage. The per variant and per animal empirical accuracy of imputation was calculated as a correlation between imputed genotypes (dosage and best-guess) and quality-filtered real genotypes of 32,614 variants (32,045 SNPs and 569 indels) genotyped for the same 293 animals using a custom XT-50K bead-chip. The 17 animals with the highest (≥ 0.99) and 17 animals with the lowest (< 0.75) imputation accuracy (per animal) were selected as two groups to check the impact of sequencing depth on imputation accuracy.

GLIMPSE2 provided the average sequence depth of all imputed chromosome chunks for each animal. In addition, BedTools (Quinlan *et al.* 2010) coverage function was used to generate the variant window sequence depth (average sequence depth of 5kb window around each variant compared). These parameters were compared for animals with high and low imputation accuracy.

RESULTS AND DISCUSSION

The mean variant and animal-based imputation accuracies are summarized in Table 1. The per variant mean imputation accuracy of skim-sequenced data was similar to previously reported imputation accuracies from standard genotyping arrays (Nguyen *et al.* 2021; Nguyen *et al.* 2024). Imputation accuracy by GLIMPSE2 in this study was better when compared to the findings of Lamb *et al,* 2023, who also used GLIMPSE but on low-coverage Oxford Nanopore sequencing data.

Table 1. Mean empirical accuracy of imputation per variant and per animal for dosage or best guess genotypes, using different chunks and effective population sizes

	Mean Accuracy							
	Chunk: 4Mb Ne:1000		Chunk: 4Mb Ne: 10,000		Chunk: 4Mb Ne: 100,000		Chunk: 18 Mb Ne=1000	
Genotypes	Variant	Animal	Variant	Animal	Variant	Animal	Variant	Animal
Dosage	0.962	0.976	0.957	0.974	0.952	0.971	0.960	0.972
Best-Guess	0.955	0.974	0.953	0.971	0.950	0.970	0.957	0.970

Consistent per variant and per animal imputation accuracies were observed when using different imputation chunk sizes (4 Mb and 18 Mb), however, imputation with 4 Mb chunk size was computationally efficient. A slight but consistent increase in imputation accuracy was observed for decreasing effective population sizes (*Ne*) which is consistent with the *Ne* being much lower in cattle. Per chromosome imputation accuracy varied a little. Variants with low minor allele frequency

(MAF) had slightly higher mean imputation accuracy than those with higher MAF (Figure 1a and b). Mean imputation accuracy for Holstein and Jersey was 0.99 and 0.95, respectively. Imputation accuracy of 22 animals (Jersey) was less than 0.9. Lower accuracy of Jersey animals may be due to the smaller number of Jersey animals in the reference population compared to the Holstein. Out of 32,614, there were 154 variants with imputation accuracy less than 0.8 (randomly spread genome-wide).

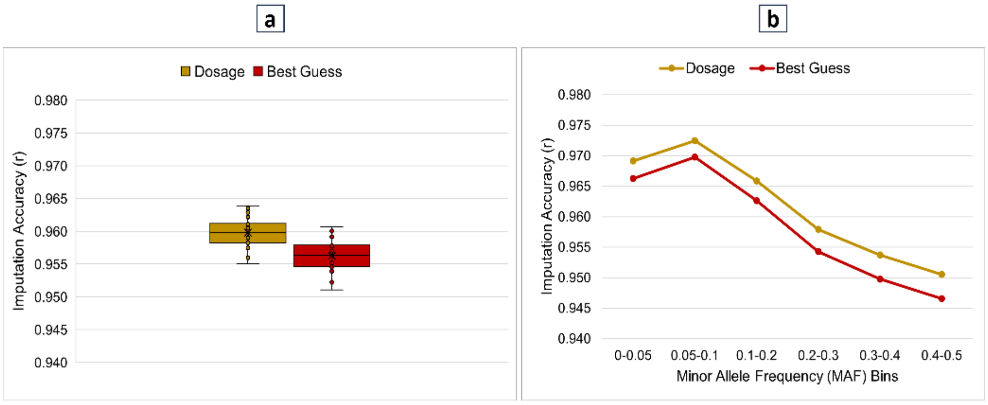

Figure 1. Mean accuracy of imputation across chromosomes and MAF bins. a: mean imputation accuracy of variants on each chromosome (best guess and dosage genotypes). **b:** mean imputation accuracy of variants grouped by MAF (best guess and dosage genotypes).

A weak correlation of 0.04 was found between the average skim sequence depth of animals and their imputation accuracy. The correlation between average sequence depths of imputation chunks (4 Mb) and imputation accuracies of animals was 0.1. The correlation between the average variant window (5 kb) sequence depth and imputation accuracy of animals was only 0.08. However, a distinct difference (p-value < 0.001) was observed in the average sequence depths of animals with the highest and lowest imputation accuracies (Figure 2a). Similarly, significant differences were observed when comparing the average imputation chunk sequencing depth (p-value < 2.2×10^{-16}) and the variant window sequence depth (p-value < 2.2×10^{-16}) between high and low imputation accuracy animals (Figure 2b and c).

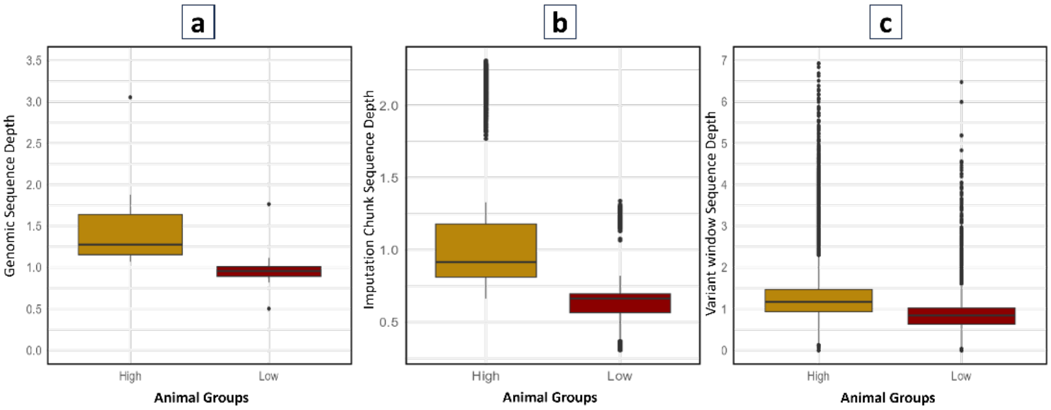

Figure 2. Comparison showing different measures of average sequence depth in animals with high or low imputation accuracy. a: genome sequence depth. **b:** sequence depth of imputation chunks. **c:** variant window sequence depth.

These differences between high and low imputation accuracy animals suggest that sequencing depth is associated with the accuracy of imputation as expected. However, this association could not be established at the population level (293 animals). It might be because of the interplay between the stochastic nature of the software algorithm and the landscape of the sequence information around each variant site within animals that impacts correct haplotyping and therefore imputation (Al Bkhetan *et al.* 2019 and Wragg *et al.* 2024). The higher average sequencing depth of high imputation accuracy animals (~1.5x) suggests that skim sequencing the samples with at least 1.5x depth can improve the overall imputation accuracy. The cost of skim sequencing assay with 1.5x depth is estimated to still be less than the cost of genotyping with a standard 50K SNP chip. Another alternative can be combining the skim sequence genotypes with "target-capture", a targeted sequencing of specific variants for higher coverage genotypes. This is of value for important SNPs such as known deleterious mutations that may be missed or have low accuracy in skim sequence genotypes.

CONCLUSION

The imputation accuracy of the skim-sequenced dataset was similar to that obtained by imputing standard SNP genotypes to WGS level. Our results suggest optimizing sequence depth versus costs. Combining skim sequencing with target capture may further improve the imputation accuracy.

REFERENCES

Al Bkhetan Z., Zobel J., Kowalczyk A., Verspoor K. and Goudey B., (2019) *BMC Bioinformatics* **20**: 1.

Browning S.R. and Browning B.L. (2007) *Am. J. Hum. Genet*. **81**: 1084.

Daetwyler H.D., Li J., Vander Jagt C.J., MacLeod I.M., Pickrell J., Vasquez M., Hoff J. and Chamberlain A.J. (2021) *Proc. Assoc. Advmt. Anim. Breed. Genet*. **24**: 78.

Davies R.W., Kucka M., Su D., Shi S., Flanagan M., Cunniff C.M., Chan Y.F. and Myers S. (2021) *Nat Genet*.**53**: 1104.

Lamb H.J., Nguyen L.T., Copley J.P., Engle B.N., Hayes B.J. and Ross E.M. (2023) *BMC Biol.* **21**: 286.

Li H. and Durbin R. (2009) *Bioinformatics*. **25**: 1754.

Li J.H., Mazur C.A., Berisa T. and Pickrell J.K. (2021) *Genome Res*. **31**: 529.

Nguyen T.V., Bolormaa S., Reich C.M., Chamberlain A.J., Medley A., Schrooten C., Daetwyler H.D. and MacLeod I.M. (2021) *Proc. Assoc. Advmt. Anim. Breed. Genet*. **24**: 267.

Nguyen T.V., Bolormaa S., Reich C.M., Chamberlain A.J., Vander Jagt C.J., Daetwyler H.D. and MacLeod I.M. (2024) *Genet. Sel. Evol*. **56**: 72.

Pasaniuc B., Rohland N., McLaren P.J., Garimella K., Zaitlen N., Li H., Gupta N., Neale B.M., Daly M.J., Sklar P. and Sullivan P.F. (2012) *Nat. Genet*. **44**: 631.

Quinlan A.R. and Hall I.M. (2010) *Bioinformatics* **26**: 841.

Rubinacci S., Hofmeister R.J., Sousa da Mota B. and Delaneau O. (2023) *Na.t Genet.* **55**; 1088.

Wragg D., Zhang W., Peterson S., Yerramilli M., Mellanby R., Schoenebeck J.J. and Clements D.N. (2024) *Genet. Sel. Evol*. **56**: 6.

Xiang R., MacLeod I.M., Daetwyler H.D., de Jong G., O'Connor E., Schrooten C., Chamberlain A.J. and Goddard M.E. (2021) *Nat. Commun*. **12**: 860.