A LINKAGE DISEQUILIBRIUM METHOD TO REPOSITION SINGLE NUCLEOTIDE POLYMORPHISM AND IMPROVE GENOTYPE IMPUTATION ACCURACY

Y. Fazel, K. Moore, S. Miller and M.H. Ferdosi

Animal Genetics Breeding Unit[*], University of New England, Armidale, NSW, 2351 Australia

SUMMARY

Incorrect positioning of single nucleotide polymorphisms (SNP) can affect imputation accuracy and decrease the accuracy of genomic prediction. This study aimed to develop a method to identify the most likely genomic position of the misplaced SNPs which have low imputation accuracy by fitting a Spline curve using linkage disequilibrium (LD) information. The accuracy of the method was validated by correctly identifying the masked position of 2,560 out of 45,918 SNP with a 100% correlation between the original and estimated positions. Candidate SNPs with low imputation accuracy (< 0.5) were assumed to be incorrectly positioned on the genome assembly. The pair-wise LD between these SNPs and other SNPs on the genome was used to fit a Spline curve. The Spline peak was considered the most likely position for the candidate SNPs. This LD-based method assigned the new position for 92% of the SNPs with low imputation accuracy and improved the mean imputation accuracy of these repositioned SNPs from 0.21 to 0.97 in Australian Brahman cattle.

INTRODUCTION

Genotype imputation can provide high-resolution genomic information that can be used to accelerate the rate of genetic gain (Piccoli *et al.* 2014). For accurate genotype imputation, a genome assembly providing the genomic positions of the SNPs is required, and for some SNPs, the reported position on the assembly may be inaccurate (Yadav *et al.* 2021). However, structural variants, genotyping errors, chip density, size of the reference population, minor allele frequency (MAF), the magnitude of LD between SNP, the number and size of the shared haplotypes between individuals in the reference and target population, genetic structure and history of the population, and SNP distribution on the genome can all have an effect on genotype imputation accuracy (Bohmanova *et al.* 2010; Lashmar *et al.* 2019; Chen *et al.* 2021). The LD information has been used to identify the correct position of the SNP in the past. Although a few studies have discussed the relationship between the LD and SNP position (Miller *et al.* 2006; Khatkar *et al.* 2010; Yadav *et al.* 2021), identifying and correcting the incorrect position of SNP on the genome is still a challenge. This study aimed to develop a new method to identify the correct position of SNPs with low imputation accuracy (< 0.5).

MATERIALS AND METHODS

Genomic data and quality control. A total number of 50,060 Australian Brahman animals, genotyped with chip densities ranging from 7k to 800k, were used to evaluate SNP imputation accuracy. Quality Control (QC) before imputation removed SNP with minor allele frequency (MAF) less than 0.05 and with a call rate of less than 0.4 across all animals. Similarly, the animals with a call rate of less than 0.2 across all SNPs in the combined data set have also been eliminated during the QC filtration (Connors *et al.* 2017). After QC, 45,973 animals and 45,918 SNPs were left for further analysis.

[*] A joint venture of NSW Department of Primary Industries and Regional Development and the University of New England

Estimation of imputation accuracy. A genotype imputation scenario from 30k to 50k with 10,000 animals in the reference population was considered the optimum scenario to evaluate the SNP imputation accuracy (Ferdosi *et al.* 2021). The population was divided randomly into 10,000 reference animals with the remaining set as the target population. Then, in the target population, SNPs were randomly masked to the lower density and imputed to the original density. The process of random division of the population and masking of the SNPs was repeated 20 times to ensure all SNPs were involved in the imputation process. FImpute version 3 (Sargolzaei *et al.* 2014) with default parameters without using the pedigree information was used for the imputation, and Pearson's correlation between observed and imputed genotypes was considered as the metric for imputation accuracy. SNPs with mean imputation accuracy above 0.99 and a standard deviation (SD) of less than 0.001 across 20 iterations were assumed to be in their correct position, in order to use them to validate the Spline method. These SNPs are hereby referred to as validation SNPs. The SNPs with a mean imputation accuracy lower than 0.5 and SD lower than 0.1 in all iterations (hereafter referred to as candidate SNPs) were identified as SNPs potentially misplaced in the genome assembly for further LD analysis and repositioning.

LD calculation. After identifying low imputation accuracy SNPs, a pair-wise LD calculation was conducted using Plink version 1.9 (Purcell *et al.* 2007) between the candidate SNPs and all other SNPs across the genome.

The Spline model. For each candidate SNP and all other SNP on each chromosome, a piece-wise cubic polynomial Spline model, specified by four coefficients, was used to capture the nonlinear relationship between LD values and positions on the genome using the Splines package, specifically the natural cubic splines (ns) function in R (Ihaka and Gentleman 1996). LD values at each genomic position were used to solve for the Spline equation coefficients. The solved polynomial equations were used to fit the Spline curve across positions on the genome. The Spline model equation is as follows;

$$X = \sum_{i=1}^{k} \beta_i \phi_i(t)$$

where X is the LD between the candidate SNP and other SNPs on the chromosome, t is the position of the SNP on the genome in base pair, β_i are coefficients estimated by the Spline model, ϕ_i is the Spline function in the interval $[t_i, t_{i+1}]$, and k is the number of intervals (the region between two consecutive knots). The coefficients of the Spline equation in each interval were calculated using the least squares fitting approach to minimise the total error between observed LD values and estimated LD values by the Spline model. The number of data points between each two consecutive knots depended on the SNP distribution across the chromosome.

Flexibility control. The flexibility (fitness) of the Spline model was controlled by defining the degree of fitness (df), which quantifies the number of intervals and the smoothness of the fitted Spline curve for each candidate SNP in each chromosome. The number of intervals also determines the model's fitness based on the number and distribution of the data points across the chromosome. Four random degrees of model fitness, based on the number of intervals (10, 20, 50, and 100), were used to find the optimum degree of fitness for the Spline model.

Peak detection. After fitting the Spline curve to identify the position of interest on the genome, the first derivative of the Spline curve equation was calculated to identify the position of the peak of the curve. Peaks and valleys are the points where the slope of the curve is zero. The highest peak across all chromosomes was considered the estimated position for each SNP candidate.

Validation of the Spline model. To test the accuracy of the Spline model in identifying the SNP position using LD information, the positions of 2,560 validation SNPs were masked and re-estimated using the Spline method. The correlation between the original and estimated positions of the validation SNPs was considered as the model accuracy.

Repositioning of the SNP with low imputation accuracy. After repositioning of the candidate SNP, a further 20 rounds of imputation were conducted to assess the imputation accuracy of the repositioned SNPs in their new positions.

RESULTS AND DISCUSSION

Figure 1 shows the accuracy of the Spline model in estimating new positions for the candidate SNPs using different degrees of fitness. The original and estimated positions of the validation SNPs had the strongest correlation when the number of intervals was the highest.

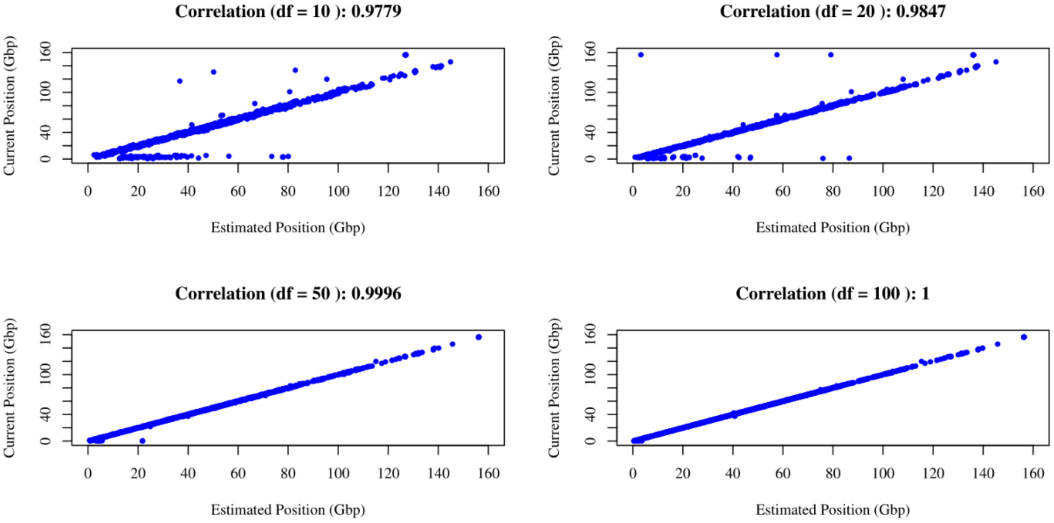

Figure 1. Validation of the Spline model using different degrees of fitness in estimating new positions using pair-wise LD information of the validation SNPs in Brahman

The Spline model was able to identify new positions for 34 out of 37 SNPs with low imputation accuracy. Figure 2 shows the imputation accuracy for these SNPs before and after repositioning.

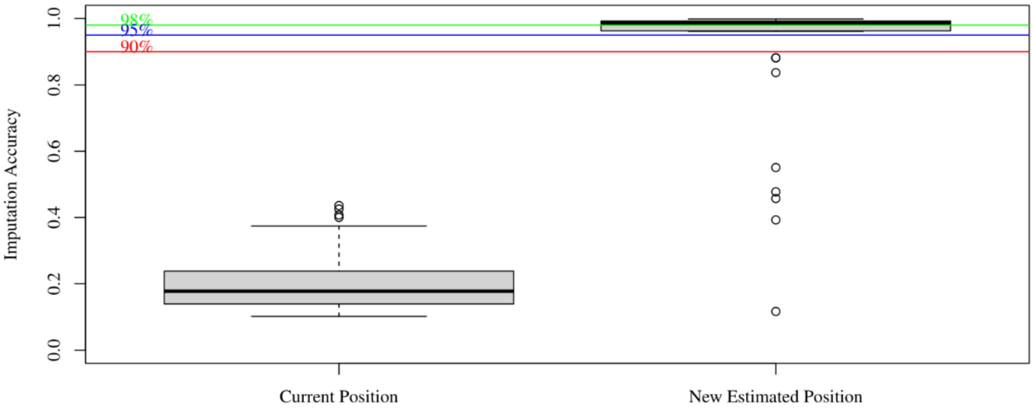

Figure 2. Imputation accuracy before and after SNP repositioning with Spline model

The Spline method was able to accurately identify the most likely position of 92% of the SNPs with low imputation accuracy on the genome using pair-wise LD information and improved the imputation accuracy of these SNPs. We observed that the incorrect position of the SNPs on the genome can contribute to lower imputation accuracy. However, the SNPs, whose new positions were assigned at the end of the chromosomes, still showed low imputation accuracy. Despite the high performance of the Spline method in identifying the correct positions, sensitivity to very low LD values and difficulties in estimating the positions of the SNPs that belong at the ends of chromosomes, can be considered as its limitations.

CONCLUSION

The incorrect position of the SNPs on the genome affects the genotype imputation accuracy. Fitting a Spline curve using pair-wise LD information can help to accurately identify the correct position of the SNPs and using newly assigned positions for the SNPs can improve the genotype imputation accuracy.

ACKNOWLEDGEMENTS

We acknowledge the Agricultural Business Research Institute (ABRI) and the Australian Brahman Breeders Association for data access. The first author was financially supported by the Animal Genetics and Breeding Unit (AGBU) scholarship.

REFERENCES

Bohmanova J., Sargolzaei M. and Schenkel F.S. (2010) *BMC Genomics* **11**: 1.

Chen L., Pryce J.E., Hayes B.J. and Daetwyler H.D. (2021) *Animals* **11**: 541.

Connors N., Cook J., Girard C., Tier B., Gore K., Johnston D. and Ferdosi M. (2017) *Proc. Assoc. Advmt. Anim. Breed. Genet* **22**: 421.

Ferdosi M., Connors N. and Khansefid M. (2021) *Proc. Assoc. Advmt. Anim. Breed. Genet* **23**: 118.

Ihaka R. and Gentleman R. (1996) *J. Comput. Graph. Stat.* **5**: 299.

Khatkar M.S., Hobbs M., Neuditschko M., Sölkner J., Nicholas F.W. and Raadsma H.W. (2010) *BMC Bioinformatics* **11**: 1.

Lashmar S., Muchadeyi F. and Visser C. (2019) *S. Afr. J. Anim. Sci.* **49**: 262.

Miller S., Hayes B., Goddard M. (2006) *Proc. World Cong. Genet. Appl. Livest. Prod.* **8**: 1.

Piccoli M.L., Braccini J., Cardoso F.F., Sargolzaei M., Larmer S.G. and Schenkel F.S. (2014) *BMC Genetics* **15**: 1.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., De Bakker P.I. and Daly M.J. (2007) *Am. J. Hum. Genet.* **81**: 559.

Sargolzaei M., Chesnais J.P. and Schenkel F.S. (2014) *BMC Genomics* **15**: 1.

Yadav S., Ross E.M., Aitken K.S., Hickey L.T., Powell O., Wei X., Voss-Fels K.P. and Hayes B.J. (2021) *BMC Genomics* **22**: 773.