Evaluating the accuracy of imputed whole-genome sequence data in admixed dairy cattle

Yu Wang[1,2], Kathryn Tiplady[1,2], Thomas J. J. Johnson[2], Chad Harland[2], Michael Keehan[1], Edwardo Reynolds[1,2], Ric G. Sherlock[2], Bevin Harris[2], Mathew D. Littlejohn[2], Richard Spelman[2], Dorian Garrick[1] and Christine Couldrey[2]

[1]AL Rae Centre of Genetics and Breeding, School of Agriculture, Massey University, Ruakura Hamilton, New Zealand

[2]Research and Development, Livestock Improvement Corporation, Hamilton, New Zealand

Background

In this study, we aimed to evaluate whether only including high-depth sequenced animals or including all sequenced animals as the reference population would benefit for whole-genome sequence level imputation under the circumstance that the number of sequenced animals is much larger compared to current effective population size. In addition, different software of phasing were used individually or combined to test for performance.

Methodology

The imputation accuracy was evaluated when high-density SNP genotypes (~644K) from an admixed population of 165,364 New Zealand dairy cattle were imputed to whole-genome sequence (WGS). WGS data were available for 336 Holstein-Friesian (H), 174 Jersey (J) and 535 H×J crossbred animals, among which 603 were sequenced with average read depth coverage >10x. The raw WGS data were aligned to the ARS-UCD1.2 bovine reference genome and variant calling was performed using GATK. VQSR filtering and standard quality control processes were conducted, resulting in ~20 million variants across the genome. Either high-depth sequenced animals or all sequenced animals were used as the reference population. The following software were singly or jointly used for phasing: Beagle 4.1, LinkPhase 3 and Beagle 5.0. Imputation was then performed from the phased data using Beagle 5.0. The quality of the imputation on chromosome 5 was evaluated by comparing the average dosage R2, or based on genotype concordance in 248 imputed animals that were subsequently sequenced for validation.

Results

Our study demonstrated that using Beagle 5.0 for phasing and imputation achieved high accuracy (average dosage R2=0.905) using all sequenced animals and the reliable variants selected from high-depth sequenced animals. It is also better compared to using all sequenced animals and directly filtered variants in all scenarios. The sequence data from 248 validation animals exhibited an error rate of 1.11%, and correlation between imputed and called variants of 98.8% (Table 1).

Table 1. Imputation accuracy (Genotype concordance, genotypic correlation and dosage R-square) between the imputed and real sequence data of 248 animals

	Nr. of variants	Genotype concordance	Genotypic correlation	Dosage R-square
Phasing method: Beagle4.1+LinkPhase3+Beagle 5.0				
$Ref_{highdepth\_filtered\_041+LP3+05}$	918,329	0.9898	0.9886	0.9127
$Ref_{all\_filtered\_041+LP3+05}$	810,898	0.9908	0.9888	0.9048
$Ref_{all\_extracted\_041+LP3+05}$	918,329	0.9898	0.9887	0.9130
Phasing method: Beagle 4.1+Beagle 5.0				
$Ref_{highdepth\_filtered\_041+05}$	918,329	0.9896	0.9884	0.9121
$Ref_{all\_filtered\_041+05}$	810,898	0.9908	0.9888	0.9048
$Ref_{all\_extracted\_041+05}$	918,329	0.9897	0.9886	0.9132
Phasing method: LinkPhase3+Beagle 5.0				
$Ref_{highdepth\_filtered\_LP3+05}$	918,329	0.9888	0.9875	0.9084
$Ref_{all\_filtered\_LP3+05}$	810,898	0.9900	0.9878	0.8977
$Ref_{all\_extracted\_LP3+05}$	918,329	0.9891	0.9879	0.9102
Phasing method: Beagle 5.0				
$Ref_{highdepth\_filtered\_05}$	918,329	0.9888	0.9875	0.9085
$Ref_{all\_filtered\_05}$	810,898	0.9900	0.9878	0.8981
$Ref_{all\_extracted\_05}$	918,329	0.9889	0.9876	0.9115

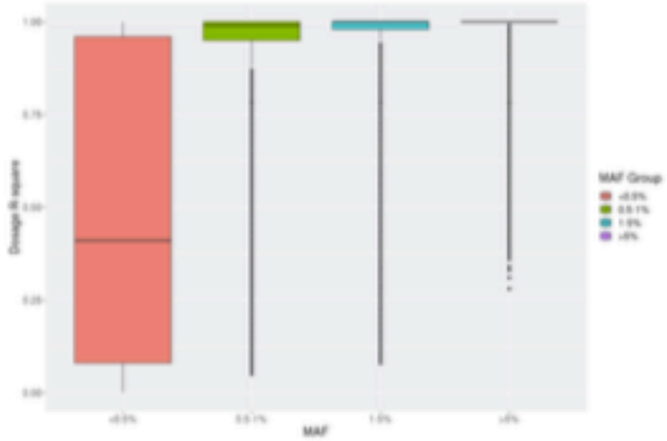

Figure 1. Dosage R^2 in different MAF categories in scenario $Ref_{all\_extracted\_05}$

Conclusion

Beagle 4.1 pre-phasing using genotype likelihood as input brought marginal benefit along with high demanding of computation resource and time, however, it may be beneficial when low-depth sequenced animals were included in the reference. The imputed dataset will be used for future genome-wide association studies for casual variant detection and genomic selection.

Acknowledgements

This study was supported by Genomics Aotearoa https://www.genomics-aotearoa.org.nz/ (project 1805). Computational resources were provided by New Zealand eScience Infrastructure (NeSI)