J. Dairy Sci. 101:9108–9127
https://doi.org/10.3168/jds.2018-14621

The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa

H. Aliloo,*[1] R. Mrode,†‡ A. M. Okeyo,† G. Ni,* M. E. Goddard,§# and J. P. Gibson*
*School of Environmental and Rural Science, University of New England, Armidale, NSW 2350, Australia
†International Livestock Research Institute (ILRI), PO Box 30709, Nairobi, Kenya
‡Scotland's Rural College, Easter Bush, Midlothian EH25 9RG, Scotland, United Kingdom
§Agriculture Victoria, AgriBio, Centre for AgriBioscience, 5 Ring Road, Bundoora, VIC 3083, Australia
#Faculty of Veterinary and Agricultural Sciences, Department of Agriculture and Food Systems, The University of Melbourne, Parkville, VIC 3010, Australia

ABSTRACT

Cost-effective high-density (HD) genotypes of livestock species can be obtained by genotyping a proportion of the population using a HD panel and the remainder using a cheaper low-density panel, and then imputing the missing genotypes that are not directly assayed in the low-density panel. The efficacy of genotype imputation can largely be affected by the structure and history of the specific target population and it should be checked before incorporating imputation in routine genotyping practices. Here, we investigated the efficacy of imputation in crossbred dairy cattle populations of East Africa using 4 different commercial single nucleotide polymorphisms (SNP) panels, 3 reference populations, and 3 imputation algorithms. We found that Minimac and a reference population, which included a mixture of crossbred and ancestral purebred animals, provided the highest imputation accuracy compared with other scenarios of imputation. The accuracies of imputation, measured as the correlation between real and imputed genotypes averaged across SNP, were around 0.76 and 0.94 for 7K and 40K SNP, respectively, when imputed up to a 770K panel. We also presented a method to maximize the imputation accuracy of low-density panels, which relies on the pairwise (co)variances between SNP and the minor allele frequency of SNP. The performance of the developed method was tested in a 5-fold cross-validation process where various densities of SNP were selected using the (co)variance method and also by alternative SNP selection methods and then imputed up to the HD panel. The (co)variance method provided the highest imputation accuracies at almost all marker densities, with accuracies being up to 0.19 higher than the random selection of SNP. The accuracies of imputation from 7K and 40K panels selected using the (co)variance method were around 0.80 and 0.94, respectively. The presented method also achieved higher accuracy of genomic prediction at lower densities of selected SNP. The squared correlation between genomic breeding values estimated using imputed genotypes and those from the real 770K HD panel was 0.95 when the accuracy of imputation was 0.64. The presented method for SNP selection is straightforward in its application and can ensure high accuracies in genotype imputation of crossbred dairy populations in East Africa.

Key words: genotype imputation, genomic selection, low-density marker panel design, East African crossbred dairy cattle

INTRODUCTION

Selection of animals based on genomic estimated breeding values (**GEBV**); that is, genomic selection (**GS**), is now a standard practice for genetic improvement of many livestock species. Genomic selection exploits the linkage disequilibrium (**LD**) between known markers and unknown causal mutations in estimation of GEBV. Genome-wide SNP are usually used as genomic markers to estimate GEBV of selection candidates that have genotypes only, based on a prediction equation that is derived from a large reference population with both genotypes and phenotypes (Meuwissen et al., 2016).

Genomic selection is especially important in situations where traditional genetic evaluations based on pedigrees are not available because of the absence of pedigree information. Smallholder dairy farmers in East Africa rear crossbred cattle to combine the adaptation features of indigenous animals with the high milk yield potential of exotic dairy breeds. These farmers do not record pedigrees and there is no current genetic evaluation to aid them in making informed breeding decisions. Based on new phenotype recording programs, GS could help East African smallholder dairy farmers to establish an effective genetic improvement program.

Received February 22, 2018.
Accepted May 26, 2018.
[1] Corresponding author: haliloo@une.edu.au

The accuracy of GEBV can increase when a larger reference population and high-density (**HD**) SNP panels are used for estimation of marker effects in the reference population. Denser panels are more effective in capturing LD between markers and QTL, and although medium density 50K assays are dense enough for reaching a useful level of LD within a breed, HD panels are required for multi-breed applications (e.g., de Roos et al., 2008). This is especially important in situations where the size of reference population for a breed is small and genotypes from other larger breeds are incorporated in genomic prediction. Although the cost of genotyping has decreased dramatically since the technology emerged, HD SNP panels are still very costly for routine use in genetic improvement of livestock species, especially in smallholder systems. A cost-effective alternative is to genotype animals with cheaper low density panels and then to infer the missing genotypes that have not been directly assayed, based on information from a reference population genotyped by an HD panel; a method called genotype imputation.

The optimal number of SNP and an appropriate algorithm for selecting them to include in a low-density array that can later be used in imputation to HD genotypes is unknown. Habier et al. (2009) suggested to genotype selection candidates using a sparse panel of evenly spaced marker across the genome and then impute the missing genotypes using co-segregation information within families. Although their proposed method could work across different breeds and traits and is independent of the genetic architecture of trait of interest, it requires availability of pedigree and HD genotypes on both parents of selection candidates. Other attempts to design low-density SNP panels has been mostly based on the use of evenly spaced markers and maximization of minor allele frequency (**MAF**) with some enrichments at chromosomal ends (e.g., Boichard et al., 2012; Bolormaa et al., 2015). Corbin et al. (2014) showed that when the low-density panels are designed to optimize equidistant spacing of markers based on LD units and to increase MAF, they can provide higher imputation accuracy and lower variations in accuracy of individual SNP than equidistant selection of SNP on base pair positions. Wu et al. (2016) described a multiple objective optimization algorithm to select SNP for low-density panels that achieved substantially higher imputation accuracies than when selecting SNP solely based on uniform distribution of map position.

Knowledge on the level and extent of LD between genome-wide markers is important because it can help to determine the required number of SNP markers for fine mapping of quantitative trait loci, GS, and genotype imputation (e.g., Sargolzaei et al., 2008; Corbin et al., 2014; Mathew et al., 2018). The structure of LD is different in different populations. It is expected that in populations with smaller effective population size (N_e) and higher average LD between markers, such as commercial dairy cattle breeds, lower number of markers will suffice. It has also been suggested that HD panels with at least 300,000 SNP are required for multi-breed applications (de Roos et al., 2008).

Existing SNP assays have been mainly designed for use in pure breeds and methods of imputation have been tested mostly in purebred populations. East African crossbred dairy cattle populations are complex admixtures of dairy *Bos taurus* breeds and indigenous African breeds. Therefore, the objectives of this study were to assess the accuracy of genotype imputation and subsequent genomic prediction in crossbred dairy cattle populations of East Africa. We compared existing arrays and methods of imputation and various methods of selecting SNP for customized arrays, including a new method based on (co)variances between SNP that are weighted by their MAF.

MATERIALS AND METHODS

Data

Population. The crossbred dairy cattle in East Africa form an admixed population resulting from many generations of crossing of African indigenous cattle to several exotic dairy breeds, mainly from Friesian, Holstein, Ayrshire and related red breeds, and Jersey. These animals are kept by smallholder dairy farmers, typically in herds of size 1 to 5 cows, and produce almost all of the milk consumed in East Africa. The majority of East African crossbred dairy cattle are bred via natural mating, with a small proportion of matings by AI to imported and locally bred purebred dairy bulls. Very few animals have pedigree records and no genetic evaluation systems or systematic breeding programs are used to aid farmers. The Dairy Genetics East Africa (**DGEA**) project collected a wide range of smallholder cow performance, animal genotype, and household data in 4 east African countries, Kenya, Uganda, Ethiopia, and Tanzania, between 2010 and 2014 to determine the needs and provide feasible solutions for short- and long-term genetic improvement of smallholder crossbred dairy cattle populations.

The genetic diversity of the crossbred cattle in relation to the indigenous breeds of the region and global reference breeds was previously presented in principal component plots by Strucken et al. (2017). They showed that the East African indigenous breeds are ancient admixtures of *Bos indicus* and African *Bos taurus* cattle where the latter is a lineage that is genetically very distinct from European *Bos taurus*. The crossbred dairy