J. Dairy Sci. 97:3128–3141
http://dx.doi.org/10.3168/jds.2013-6826

Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds

S. G. Larmer,*[1] M. Sargolzaei,*† and F. S. Schenkel*
*Centre for Genetic Improvement of Livestock, Animal and Poultry Science Department, University of Guelph, Guelph, Ontario, N1G 2W1 Canada
†L'Alliance Boviteq, Saint-Hyacinthe, Quebec, J2T 5H1, Canada

ABSTRACT

Genomic selection requires a large reference population to accurately estimate single nucleotide polymorphism (SNP) effects. In some Canadian dairy breeds, the available reference populations are not large enough for accurate estimation of SNP effects for traits of interest. If marker phase is highly consistent across multiple breeds, it is theoretically possible to increase the accuracy of genomic prediction for one or all breeds by pooling several breeds into a common reference population. This study investigated the extent of linkage disequilibrium (LD) in 5 major dairy breeds using a 50,000 (50K) SNP panel and 3 of the same breeds using the 777,000 (777K) SNP panel. Correlation of pair-wise SNP phase was also investigated on both panels. The level of LD was measured using the squared correlation of alleles at 2 loci (r^2), and the consistency of SNP gametic phases was correlated using the signed square root of these values. Because of the high cost of the 777K panel, the accuracy of imputation from lower density marker panels [6,000 (6K) or 50K] was examined both within breed and using a multi-breed reference population in Holstein, Ayrshire, and Guernsey. Imputation was carried out using FImpute V2.2 and Beagle 3.3.2 software. Imputation accuracies were then calculated as both the proportion of correct SNP filled in (concordance rate) and allelic R^2. Computation time was also explored to determine the efficiency of the different algorithms for imputation. Analysis showed that LD values >0.2 were found in all breeds at distances at or shorter than the average adjacent pair-wise distance between SNP on the 50K panel. Correlations of r-values, however, did not reach high levels (<0.9) at these distances. High correlation values of SNP phase between breeds were observed (>0.94) when the average pair-wise distances using the 777K SNP panel were examined. High concordance rate (0.968–0.995) and allelic R^2 (0.946–0.991) were found for all breeds when imputation was carried out with FImpute from 50K to 777K. Imputation accuracy for Guernsey and Ayrshire was slightly lower when using the imputation method in Beagle. Computing time was significantly greater when using Beagle software, with all comparable procedures being 9 to 13 times less efficient, in terms of time, compared with FImpute. These findings suggest that use of a multi-breed reference population might increase prediction accuracy using the 777K SNP panel and that 777K genotypes can be efficiently and effectively imputed using the lower density 50K SNP panel.
Key words: linkage disequilibrium, dairy, imputation

INTRODUCTION

The advent of genomic selection has been a major breakthrough in the breeding programs of many dairy breeds, and it continues to improve with the advent of new technologies to increase the accuracy of selection and the reliability of the methods used. To begin genomic selection within a breed or group of breeds, an important step is assessing the level of linkage disequilibrium (**LD**; Meuwissen et al., 2001). Linkage disequilibrium is a measure of the nonrandom association of alleles that helps us to infer the alleles present at other loci, especially at QTL that have an effect on phenotypes of interest. Numerous studies have found high levels of LD between adjacent marker pairs as well as LD extending over tens of centimorgans (Farnir et al., 2000). However, useful LD in Holsteins was only found at distances <100 kb (Sargolzaei et al., 2008). Useful LD was defined as an r^2 value >0.3, which is the level of LD deemed necessary for successful association studies (Ardlie et al., 2002). Calus et al. (2008) showed, however, that accuracy of genomic selection can be implemented accurately when the level of LD, based on r^2, is >0.2.

The 2 factors that affect accuracy of genomic selection that can be controlled are the level of LD between markers and QTL and the size of the reference population of animals used to estimate SNP effects on phenotypes (Hayes et al., 2009). The level of LD between markers and QTL can be controlled by utilizing very dense mark-

Received March 18, 2013.
Accepted January 14, 2014.
[1]Corresponding author: slarmer@uoguelph.ca

er panels. Genomic selection, as has been implemented in the Holstein, Jersey, and Brown Swiss populations in North America, uses the 50,000 marker (**50K**) panel. With the advent of SNP panels with up to 777,000 markers (**777K**), the accuracy of genomic selection can be increased. Calus et al. (2008) found, using simulated data, that if r^2 is increased from 0.1 to 0.2, the accuracy of genomic selection increased from 0.68 to 0.82. Brito et al. (2011) reported that for similar accuracy of genomic selection, approximately half as many animals would be needed in the training population when high-density panels were used. Erbe et al. (2012) showed slight gains in genomic EBV (**GEBV**) accuracy when using a high-density marker panel (~800,000) compared with using the 50K panel in Holstein and Jersey populations. This increase in accuracy is relatively small because enough linkage within a breed to effectively implement genomic selection can be found on the 50K panel, especially with large reference populations of genotyped animals, such as is available in the Holstein breed. For a population with fewer animals genotyped (n = 540), Erbe et al. (2012) also showed that a combined reference population could increase GEBV accuracy by an average of 4% for various production traits when using a high-density marker panel.

These dense marker panels, however, are expensive and can be a major limitation to the number of animals genotyped. The advent of imputation has significantly aided this problem, making it possible for many animals to be genotyped with lower-density marker panels and imputed to higher density for genomic selection. In Canada, imputation takes place regularly from approximately 3,000 (**3K**) and 6,000 (**6K**) SNP to the 50K SNP panel by using the FImpute program (Sargolzaei et al., 2011a) and both the family and population algorithms. For a large reference population, imputation from ~3K to ~50K has been shown to be accurate using both family-based and population-based methods, with accuracy exceeding 0.90. When family information is available, family-based methods do have an advantage over population-based methods (Zhang and Druet, 2010).

Once the level of LD between marker and causative mutation has been maximized by using the highest density panel available or with full sequence, the reference population must be grown to increase the accuracy of genomic prediction. Luan et al. (2009) showed an increase in accuracy of approximately 5% for genomic selection of production traits in Norwegian Red cattle when the reference population grew from 250 to 400 animals. Liu et al. (2011) also saw an increase in the variance of SNP effects as the reference population size grew, allowing for larger SNP effects to be detected and selected upon. One way to increase the training population size for genomic selection would be to use a breed with more genotyped individuals in the reference population for a breed with fewer genotyped animals. It was found that for Holsteins and Jerseys, one breed could not be directly used as the sole training population for selection in the other breed using the 50K marker panel, but results were more promising when both breeds were combined into a common population, although gains were modest (Pryce et al., 2011). For genomic selection to be effective, markers and QTL need to be in the same linkage phase across the populations, and linkage phase must be consistent from the training population to the validation population. de Roos et al. (2008) estimated that for breeds to be pooled into a common reference population that increases accuracy of genomic prediction, approximately 300,000 evenly spaced markers would be needed to ensure adequate consistency of phase across breeds in the multi-breed reference population for genomic selection.

The goal of this study was to explore the extent of LD in 5 Canadian dairy breeds using the 50K SNP panel, as well as that in 3 breeds using the 777K SNP panel. Consistency of phase was then explored to determine the possible use of a multi-breed reference population for genomic selection in Ayrshires and Guernseys using both the 50K and 777K SNP panels. Accuracy of imputation from 6K or 50K to 777K using real Ayrshire, Guernsey, and Holstein data was then calculated. Both a population-based (Beagle 3.3.2; Browning and Browning, 2007) and family and population-based (FImpute v2.2; Sargolzaei et al., 2011a) method were explored. The effects of a multi-breed reference population and the effect of having direct ancestors genotyped were also examined. If phase consistency is high, multi-breed imputation is expected to increase accuracy for populations with few genotyped individuals. In addition, imputation from 6K to 777K using a 2-step approach was assessed in the Holstein breed.

MATERIALS AND METHODS

Data

Genotypes from Holstein (n = 47,433) Jersey (n = 4,517), and Brown Swiss (n = 1,566) were taken on animals from the North American Collaboration on Genomic Prediction, many of which were proven sires. Genotyping was performed using the Illumina 50K BeadChip (Illumina Inc., San Diego, CA). In addition, 476 proven Ayrshire bulls and 61 proven Guernsey bulls were genotyped with the BovineHD (777K) panel (Illumina Inc.). Holstein 777K genotypes (n = 1,115) were also obtained from The North American Collaboration on Genomic Prediction.