Abstract W91

Accuracy of Genomic Imputation in a Thai Multibreed Dairy Cattle Population

Danai Jattawa[1,2], Skorn Koonawootrittriron[1], Mauricio A. Elzo[2] and Thanathip Suwanasopee[1]

Kasetsart University, Chatuchak, Bangkok 10900, Thailand[1]

University of Florida, Gainesville, FL 32611-0910, USA[2]

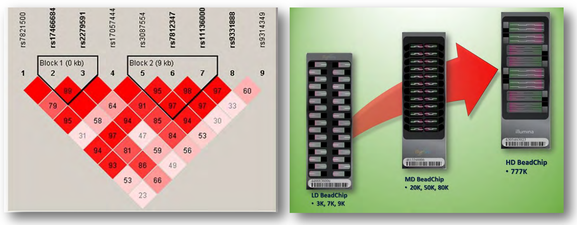
SUMMARY

The objective of this study was to investigate the accuracy of imputation from low (**LDC**) to moderate density SNP chips (**MDC**) in a Thai Holstein-Other multibreed dairy cattle population. Dairy cows with complete pedigree information (n = 1,110) from 129 dairy farms were genotyped with GeneSeek GGP20K (n = 570) and GGP26K (n = 540) BeadChips. After checking for genotypic quality, 16,387 SNP in common between the GGP20K and GGP26K were used to represent **MDC** in this study. Cows were divided into two groups, a reference group (n = 778) and a test group (n = 332). The SNP genotypes chosen for the test group were those SNP located in positions corresponding to GeneSeek GGP9K (n = 7,356). The **LDC** to **MDC** genomic imputation was carried out using three different methods, namely a population-based algorithm in the Beagle software (**PBG**), a population-based algorithm in the FImpute software (**PFI**), and a combined family and population-based algorithm in FImpute (**CFI**). Imputation accuracies within and across chromosomes were calculated as ratios of correctly imputed genotypes to overall imputed genotypes. Imputation accuracy for the three methods ranged from 76.31% to 93.91%. The **CFI** had slightly higher imputation accuracy (93.91%) than **PFI** (93.56%) and both methods were substantially more accurate than PBG (76.31%). Noticeably most chromosomes that showed either high or low imputation accuracies were the same chromosomes that had high and low average linkage disequilibrium (defined here as the correlation between pairs of adjacent SNP within chromosomes less than 5 MB apart). This suggested that choosing sets of SNP with high levels of average linkage disequilibrium would improve imputation accuracy. ***Results clearly indicated that FImpute software (population or combined family-population) were more suitable than Beagle for genotype imputation in this Thai multibreed population. Perhaps additional increments in imputation accuracy could be achieved by discarding SNP with low levels of average LD, and by increasing the completeness of pedigree information.***

INTRODUCTION

Genomic information has been widely used in livestock animal studies and has come to play an important role in characterization and evaluation of dairy cattle. Genomic selection utilizes genomic information to increase the rate of genetic progress in dairy populations. High density genotyping data increases the effectiveness of genomic selection, but genotyping costs may be prohibitively high in some cases. An alternative is to use imputation where animals genotyped with a lower density chip are imputed to a higher density chip using information from animals genotyped with the higher density chip. This approach could drastically reduce genotyping costs compared to using only high density chips. Imputation in dairy cattle is regularly utilized in many countries, and several software packages have been developed for this purpose. These software packages have yielded accuracies of imputation ranging from 88% to 99% depending on the software and the dairy population. Imputation accuracy of software packages was primarily tested on purebred dairy populations and under temperate conditions. However, the Thai dairy cattle population is multibreed (animals can be composed of fractions of up to 8 different breeds) and they are under tropical conditions. ***Thus, the objective of this study was to investigate the accuracy of imputation from low (9K) to moderate SNP density (20K) in a Thai Holstein-Other multibreed dairy cattle population genotyped with GeneSeek GGP20K and GGP26K BeadChips.***

Thai Holstein-Other multibreed dairy cattle composed of fractions from up to 8 different breeds



MATERIALS AND METHODS

Animals and genotypes

Genotypic information from moderate density chips (**MDC**; GGP20K and GGP26K; GeneSeek, Lincoln, NE, USA) was available on 1,110 Thai dairy cows (570 cows genotyped with GGP20K and 540 cows genotyped with GGP26K). SNP on the Y, X, and mitochondrial chromosomes, SNP of unknown chromosome position, SNP with a MAF lower than 0.01, and SNP with a call rate lower than 0.9 were eliminated. After these quality checks, 16,387 SNP in common between GGP20K and GGP26K were kept to represent **MDC**. SNP genotypes were located in positions described by the UMD 3.1 genome build (University of Maryland, College Park, MD, USA).

The pedigree file contained information from 2,283 animals. All cows were crossbreds (composed of fractions of up to 8 different breeds; Holstein, Brahman, Jersey, Brown Swiss, Red Dane, Red Sindhi, Sahiwal and Thai Native breed). Cows were from 129 farms and had their first calving between 2004 and 2014.

Genomic Analyses

Linkage disequilibria (**LD**) among GGP20K SNP genotypes in the Thai dairy cattle population were computed as correlation coefficients (r^2) between pairs of adjacent SNP within chromosomes that were less than 5 Mb apart. The r^2 values were estimated using Haploview 4.2.

Cows were divided into two groups: a reference group and a test group. Sorting by year of birth was carried out to identify as many dams with strong genetic relationships as possible to be included in the reference group. With this process, 778 cows born before 2010 were included in the reference group, and the remaining 332 cows (born in or after 2010) became part of the test group. The SNP genotypes in the test group were the subset of SNP in common between the GGP20K and GGP26K that were also represented in the GGP9K BeadChip (**LDC**: n = 7,380 SNP).

The **LDC** to **MDC** genomic imputation was performed using three different methods, namely the population-based algorithm in Beagle 3.3 (**PBG**), a population-based algorithm in FImpute 2 (**PFI**), and a combined family-and-population-based algorithm in FImpute 2 (**CFI**). Beagle and Fimpute were chosen in this study because these software packages were found to have high imputation accuracy in livestock populations (Johnston *et al.,* 2011; Sun *et al.,* 2012; Larmer *et al.,* 2014). After completing the imputation process, imputed genotypes were compared to the true genotypes, and imputation accuracy was evaluated within and across chromosomes. Imputation accuracy was computed using the expression:

$$\text{Imputation Accuracy} = \frac{\text{Correctly Imputed SNP}}{\text{Overall Imputed SNP}} \times 100$$

Table 1. Imputation accuracy from LDC to MDC using Beagle (population-based; PBG), FImpute (population-based; PFI), and FImpute (combined family &population-based ; CFI) in a Thai multibreed dairy cattle population

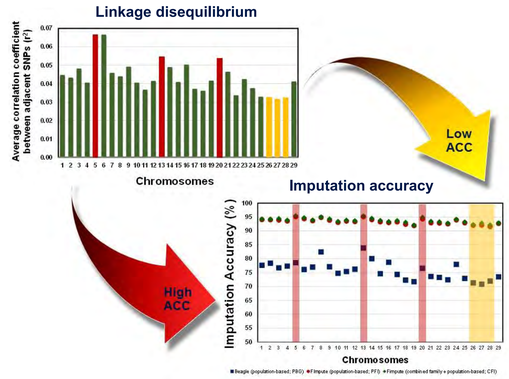
Software	Algorithm	Imputed SNP	Correct SNP	Accuracy (%)
Beagle	Population-based (PBG)	2,940,584	2,243,834	76.31
FImpute	Population-based (PFI)	2,940,584	2,751,091	93.56
	Family + Population-based (CFI)	2,940,584	2,761,585	93.91

RESULTS AND DISCUSSION

The average **LD** estimated for **MDC** in this Thai multibreed dairy population ranged from 0.03 (chromosome 27) to 0.07 (chromosome 6). The five highest average **LD** were in chromosomes 5, 6, 13, 16, and 20, and the five lowest average **LD** occurred in chromosomes 22, 25, 26, 27 and 28. Imputation accuracy from **LDC** to **MDC** for the three methods ranged from 76.31% to 93.91% (**Table 1**). The **CFI** had slightly higher imputation accuracy (93.91%) than **PFI** (93.56%) and both methods were substantially more accurate than **PBG** (76.31%).

Several studies have reported different results in terms of the variation in imputation accuracy obtained between FImpute and Beagle. Johnston *et al.* (2011) reported that FImpute outperformed Beagle for genomic imputation from 3K to 50K in Brown Swiss (0.7%) and Holstein (1.6%) populations. Similarly, FImpute slightly outperformed Beagle for imputation from 50K to 777K in Guernsey (0.2%) and Ayrshire (0.2%) populations (Larmer *et al.,* 2014). Conversely, Beagle performed better than FImpute for imputation from 7K to 50K in an Angus population (1.16%; Sun, *et al.,* 2012), and also from 3K to 50K in Swedish and Finnish Red cattle (1.1%; Ma *et al.,* 2013). These different reports suggests that imputation accuracy depends on compatibility between algorithm (software) and population structure. ***Thus, results here indicated that FImpute was more compatible with the Thai dairy cattle population than Beagle.***

This study showed that both algorithms of the FImpute software package performed acceptably (more than 93% correctly) in the Thai dairy cattle population. However, there was a slight difference in imputation accuracy between the use of combined family-population and population only algorithms. The gain in imputation accuracy was found to be only 0.35% when using combined algorithm compared with the population only algorithm. The small gain in imputation accuracy may have been due to the small number of genotyped parents and older ancestors included in the reference group. Only 32 dams of imputed animals (4% of animals in the reference group) were used in this group, which reflects a low relationship between the reference and target groups. Several studies have reported that the imputation accuracy increases with the increasing relationship between these two groups, particularly when imputing from low to moderate density (e.g. Zhang and Druet, 2010; Ma *et al.,* 2013; Carvalheiro *et al.,* 2014). High genetic relationships decreased imputation error rates. Further, Sargolzaei *et al.* (2010) also reported that family-based was the most effective algorithm in a dairy population where high level of pedigree information and a large reference population were available. When pedigree information is not sufficient, FImpute assumes that the remaining animals are unrelated and impute them using a population-based algorithm (Lamer *et al.,* 2014). ***Increasing the number of genotyped ancestors with close genetic relationships in the reference and the target groups is expected to increase the imputation accuracy in the Thai dairy cattle population.***



Imputation accuracy was also computed for each of the 29 autosomal chromosomes from the 1,110 Thai dairy cattle. Accuracies were similar and consistent across chromosomes when imputed using FImpute, for both combined family-population and population-based algorithms, but were variable for the Beagle software. Among the 29 chromosomes, imputation accuracies ranged from 92.05% to 95.46% for **CFI**, 91.57% to 95.12% for **PFI**, and 70.85% to 83.84% for **PBG**. High and low levels of imputation accuracy were obtained for each method in the same sets of chromosomes. Chromosomes 5, 8, 13, 14, and 20 had high imputation accuracy, whereas chromosomes 18, 19, 26, 27 and 28 showed low imputation accuracy.

Levels of imputation accuracy were likely affected by the level of **LD** in the reference group. This occurred with Beagle and FImpute, where sets of high and low imputation accuracies occurred in the same chromosomes. In addition, most chromosomes within the set of high imputation accuracy exhibited high average LD, but most chromosomes within the set of low imputation accuracy had low average LD. ***These results suggested that imputation accuracy could be improved by removing SNP with very low level of average LD before performing imputation.***

FINAL REMARKS

- ***FImpute had higher imputation accuracy than Beagle in this Thai Multibreed Dairy Population***
- ***Family-Population was slightly better than population based imputation***
- ***Imputation accuracy could be increased by:***
  - ***Increasing number of genotyped related animals in the reference and target populations***
  - ***Removing SNP with low average LD***

LITERATURE CITED

Carvalheiro, R., S.A. Boison, H.H.R. Neves, M. Sargolzaei, F.S. Schenkwl, Y.T. Utsunomiya, A.M.P. O'Brien, J. Sölkner, J.C. McEwan, C.P. vanTassell, T.S. Sonstegard and J. F. Garcia. 2014. Genet. Sel. Evol. 46: 69.

Johnston, J., G. Kistemaker and P.G. Sullivan. 2011. Comparison of different imputation methods, Interbull Bulletin No.44. Stavanger, Norway.

Larmer, S.G., M. Sargolzaei and F.S. Schenkel. 2014. J. Dairy Sci. 97: 3128-3141.

Ma, P., R.F. Brøndum, Q. Zhang, M.S. Lund and G. Su. 2013. J. Dairy. Sci. 96: 446-467.

Sargolzaei, M., J.R. Chesnais and F.S. Sschenkel. 2010. GEB Open Industry Session, Saint-Hyacynthe, Quebec, Canada.

Sun, C., X-L Wu, K.A. Weigel, G.J.M. Rosa, S. Bauck, B.W. Woodward, R.D. Schnabell, J.F. Taylor and D. Gianola. 2012. Gent. Res. Camb. 94: 133-150.

Zhang, Z. and T. Druet. 2010. J. Dairy Sci. 93 (11): 5487-5494.