Livestock Science 180 (2015) 27–33

Contents lists available at ScienceDirect

Livestock Science

journal homepage: www.elsevier.com/locate/livsci

Genome-wide linkage disequilibrium in a Thai multibreed dairy cattle population

Thawee Laodim [a], Skorn Koonawootrittriron [a,*], Mauricio A. Elzo [b], Thanathip Suwanasopee [a]

[a] Department of Animal Science, Kasetsart University, Bangkok 10900, Thailand
[b] Department of Animal Sciences, University of Florida, Gainesville, FL 32611, USA

ARTICLE INFO

Article history:
Received 6 February 2015
Received in revised form
25 June 2015
Accepted 30 June 2015

Keywords:
Dairy cattle
Linkage disequilibrium
Single nucleotide polymorphism
Tropical regions

ABSTRACT

The level of linkage disequilibrium (LD) plays an important role in increasing the power to detect associations for mapping quantitative trait loci in the genome and in increasing the accuracy of prediction of genomic estimated breeding values (GEBV). Thus, the objectives of this study were to evaluate the extent of LD in Thai multibreed dairy cattle and to determine factors that influence the estimation of LD. A total of 1413 multibreed dairy cows were genotyped for 8220 SNPs, covering 2507.24 Mb of the genome. The mean of minor allele frequencies (MAF) across autosomes was 0.37. All possible SNP pairs on the same chromosome were used to estimate LD across the 29 autosomes. High levels of LD were found in autosomes, particularly between SNP pairs at distances shorter than 50 kb. The mean of D' (linkage disequilibrium relative to its maximum) and r^2 (coefficient of correlation squared) for SNPs at 40–50 kb apart were 0.694 and 0.202, respectively. Overestimation of D' occurred when the MAF threshold was low (0.05). The r^2 was high when the MAF threshold was higher than 0.20, especially when the distance between markers was shorter than 50 kb. The minimum sample sizes required to obtain accurate measures of LD were 177 for D' and 89 for r^2. Results from this research will be useful for genome-wide association studies and genomic selection of dairy cattle in tropical regions.

1. Introduction

Dairy cattle in Thailand and other tropical countries are largely multibreed. The vast majority of cattle in the Thai multibreed dairy population are crossbred (91%). Their genetic composition is usually over 75% Holstein (H) and the remainder comes from various *Bos indicus* (e.g., Red Sindhi, Sahiwal, Brahman and Thai Native) and *Bos taurus* (e.g., Brown Swiss, Jersey and Red Danish) breeds. An animal could have as many as eight different cattle breeds represented in it (Koonawootrittriron et al., 2009). For this reason, Thai multibreed dairy cattle populations are different from cattle populations in other countries. Genetic evaluation programs for economically important traits of Thai multibreed dairy cattle currently use a multibreed animal model based on level of *H* fraction of the animals. The main focus of these programs is on milk yield, the primary selection criterion for dairy genetic improvement by Thai dairy farmers.

An efficient alternative to improve the accuracy of selection and to speed up genetic progress for this trait could be genomic selection. Genomic selection refers to selection based on genomic breeding values (GEBV) of animals computed using prediction equations that utilize a large number of markers (Meuwissen et al., 2001; Solberg et al., 2008). The accuracy of GEBV depends on the level of linkage disequilibrium (LD) between markers and quantitative trait loci (QTL; Hayes et al., 2009). The LD refers to nonrandom associations between alleles at two loci and plays a fundamental role in gene mapping for economically important traits (Reich et al., 2001) and in genome-wide association studies (Yang et al., 2014).

The LD is also of interest for what it reveals about history because the distribution of LD is determined in some of the genome regions by the population history (McKay et al., 2007). In addition, studies of LD may enable a better understanding of the biology of recombination (Ardlie et al., 2002) because it is difficult to use pedigree to estimate the rate of homologous gene conversion or variation in recombination rates at very short distances due to very low rates of occurrence of these events (Pritchard and Przeworski, 2001).

The level of LD between markers and QTL can be quantified with the two most common parameters D' and r^2 (Khatkar et al., 2008; Bohmanova et al., 2010; Espigolan et al., 2013). Both

* Corresponding author. Fax: +662 579 1120.
E-mail address: agrskk@ku.ac.th (S. Koonawootrittriron).

http://dx.doi.org/10.1016/j.livsci.2015.06.021
1871-1413/

parameters range from 0 (incomplete disequilibrium) to 1 (complete disequilibrium), but their interpretation are slightly different. A value of $D'=1$ indicates that two SNPs have not been separated by recombination, recurrent mutation and gene conversion during the history of the sample. Conversely, $D'<1$ indicates the complete disruption of ancestral LD, and its relative magnitude cannot be interpreted. Estimates of D' are strongly inflated in small samples and SNPs with low allele frequencies. Therefore, D' values near 1 are not useful for comparisons of the strength of LD between studies, or for measuring the extent of LD (Ardlie et al., 2002). An r^2 value represents a statistical correlation between two sites and takes the value of 1 only when two SNPs have not been separated by recombination and when the markers also have the same allele frequencies (Pritchard and Przeworski, 2001). Hence, r^2 is preferred for measuring of LD in the context of association mapping because there is a simple inverse relationship between r^2 and the sample size required to detect association between SNPs (Pritchard and Przeworski 2001; Ardlie et al., 2002).

Previous studies on LD in dairy cattle were based on high density of SNPs at short distances in purebred cattle under temperate conditions (Sargolzaei et al., 2008; Bohmanova et al., 2010; Espigolan et al., 2013). Khatkar et al., (2008) reported that $r^2 \geq 0.2$ was observed for SNPs less than 40 kb apart in an Australian Holstein–Friesian population. Similarly, a level of $r^2 \geq 0.2$ in North American Holstein was observed at distances between markers up to 60 kb (Bohmanova et al., 2010). A level of $r^2 \geq 0.2$ was observed at a distance of 75 kb between SNPs in German Holstein cattle by Qanbari et al. (2010). Variation in the extent of LD depends on factors such as population structure, natural selection, and variable recombination rates (Ardlie et al., 2002). The LD could also differ between purebred and multibreed dairy populations as a results of different allele frequencies in the parental breeds (Varoneze et al., 2014). Thus, the objective of this research was to evaluate LD and describe the extent and pattern of LD on autosomes under four minor allele frequency and seven sample size scenarios in a Thai multibreed dairy cattle population using 8220 SNPs.

2. Material and methods

2.1. Animals and data

Animals (1413 cows) in this study were members of the Thai multibreed dairy cattle population, which was described by Koonawootrittriron et al. (2009). Breeds present in this population were Holstein (H), Jersey, Red Danish, Brahman, Red Sindhi, Sahiwal, Thai Native, and other breeds. Nearly all cows in this population were crossbred (97%), and the breed composition of an animal could include fractions from up to seven different breeds. Holstein fractions in crossbred animals ranged from 28% to 99%. Cows were reared by farmers (195 farms) in three regions of the country (901 cows from 78 farms in Central Thailand; 298 cows from 67 farms in Southern Thailand; 214 cows from 50 farms in Northeastern Thailand). Cows were born between 1997 and 2011 and had complete pedigree and first lactation information.

2.2. Blood samples and genotypes

Blood samples were taken from the caudal vein (9 ml), kept below 4 °C, and then transported from the farm to the laboratory at Kasetsart University in Bangkok within 24 h. The DNA from each sample was extracted and purified by applying a protocol of the MasterPure™ DNA Purification Kit (Epicentre®, USA). The quantity of DNA per sample was measured using a NanoDrop 2000 (Thermo Fisher Scientific Inc., Wilmington, DE, USA). The DNA was accepted as pure when the purity ratio is 260/280 of approximately 1.8, and the DNA concentration was higher than 15 ng/μl.

The SNP genotyping was done by GeneSeek Inc. (Lincoln, NE, USA) using the GeneSeek Genomic Profiler low density (GGP-LD) BeadChip that utilizes the Illumina Infinium® chemistry (Illumina, San Diego, CA, USA). Each chip contains a total of 8810 SNPs of which 8305 SNP loci had known physical locations on the 29 autosomes (sex chromosomes were ignored in this study). The SNPs with minor allele frequency (MAF) of less than 0.05 were filtered out. After filtering, a total of 8220 SNPs loci were included in the final analysis.

2.3. Measures of linkage disequilibrium

Linkage disequilibrium (LD) is a measure of the non-random association between two alleles that helps to infer the alleles at QTL that influence phenotypes of interest. Currently, the most commonly used parameters to measure LD are D' and r^2 (Zhao et al., 2005). The D' is a measure of LD relative to the maximum possible value given the allele frequency of SNPs. The D' was considered from the frequencies of the haplotype of the SNP pairs, and it was calculated as follows:

$$D' = \begin{cases} \frac{D}{\min\,(f(A)\times f(b), f(a)\times f(B))} \text{ if } D>0 \\ \frac{D}{\min\,(f(A)\times f(B), f(a)\times f(b))} \text{ if } D<0 \end{cases}$$

and

$$D = f(AB)\times f(ab) - f(Ab)\times f(aB),$$

where $f(A)$, $f(a)$, $f(B)$ and $f(b)$ denote the allele frequencies of SNPs, and $f(AB)$, $f(Ab)$, $f(aB)$ and $f(ab)$ are the four haplotype frequencies in the population (Lewontin, 1964).

The r^2 is the square of correlation between pairs of SNP. This parameter can be used as a standardization measurement of LD between alleles of two loci (Zhao et al., 2005). The r^2 is generally less inflated in small samples than D' (Ardlie et al., 2002). This measure can be calculated from D and allele frequencies of the SNPs following Hill and Robertson (1968).

$$r^2 = \frac{(D)^2}{f(A)\times f(a)\times f(B)\times f(b)}$$

The D' and r^2 for all pair-wise combinations of the SNPs on each autosome were inserted into software Haploview (Barrett et al., 2005) to verify SNP quality after excluding the SNPs with MAF <0.05 and Hardy–Weinberg equilibrium with $P<0.0001$. To compare LD over autosomes, the maximum distance between SNP pairs was limited to 5 Mb.

2.4. Effect of MAF and sample size on linkage disequilibrium

The effect of MAF on estimates of D' and r^2 was evaluated using four different minimum MAF thresholds (0.05, 0.10, 0.15 and 0.20). Because LD decays as physical distance between loci increases, SNPs were classified into three groups based on distance between loci (every 10 kb, 100 kb and 1 Mb; 23 groups in total). Then, D' and r^2 were estimated for each MAF threshold by distance between loci combination to assess LD variation in this population.

To examine the effect of sample size on estimated values of D' and r^2, seven sample sizes were considered: (1) 45 cows (1/32 or 3.125%); (2) 89 cows (1/16 or 6.25%); (3) 177 cows (1/8 or 12.5%); (4) 354 cows (1/4 or 25%); (5) 707 cows (1/2 or 50%); (6) 1059 cows (3/4 or 75%) and (7) 1413 cows (1 or 100%). Cows for samples 1–6 were randomly drawn from the full dataset by taking bootstrap subsamples with replacement (Teare et al., 2002).

Average values of D' and r^2 were calculated for each pair of SNPs at the specified distance ranges in each sample size. The SNPs