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

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 1,413 multibreed dairy cows were genotyped for 8,220 SNPs, covering 2,507.24 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 to 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.

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

Keywords dairy cattle, linkage disequilibrium, single nucleotide polymorphism, 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 non-random 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).

48 The LD is also of interest for what it reveals about history because the distribution of
49 LD is determined in some of the genome regions by the population history (McKay et al.,
50 2007). In addition, studies of LD may enable a better understanding of the biology of
51 recombination (Ardlie et al., 2002) because it is difficult to use pedigree to estimate the rate
52 of homologous gene conversion or variation in recombination rates at very short distances
53 due to very low rates of occurrence of these events (Pritchard and Przeworski, 2001).
54 The level of LD between markers and QTL can be quantified with the two most
55 common parameters D′ and r^2 (Khatkar et al., 2008; Bohmanova et al., 2010; Espigolan et al.,
56 2013). Both parameters range from 0 (incomplete disequilibrium) to 1 (complete
57 disequilibrium), but their interpretation are slightly different. A value of $D' = 1$ indicates that
58 two SNPs have not been separated by recombination, recurrent mutation and gene conversion
59 during the history of the sample. Conversely, $D' < 1$ indicates the complete disruption of
60 ancestral LD, and its relative magnitude cannot be interpreted. Estimates of D′ are strongly
61 inflated in small samples and SNPs with low allele frequencies. Therefore, D′ values near 1
62 are not useful for comparisons of the strength of LD between studies, or for measuring the
63 extent of LD (Ardlie et al., 2002). An r^2 value represents a statistical correlation between two
64 sites and takes the value of 1 only when two SNPs have not been separated by recombination
65 and when the markers also have the same allele frequencies (Pritchard and Przeworski,
66 2001). Hence, r^2 is preferred for measuring of LD in the context of association mapping
67 because there is a simple inverse relationship between r^2 and the sample size required to
68 detect association between SNPs (Pritchard and Przeworski 2001; Ardlie et al., 2002).
69 Previous studies on LD in dairy cattle were based on high density of SNPs at short
70 distances in purebred cattle under temperate conditions (Sargolzaei et al., 2008; Bohmanova
71 et al., 2010; Espigolan et al., 2013). Khatkar et al., (2008) reported that $r^2 \geq 0.2$ was observed
72 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 (Veroneze 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 8,220 SNPs.

2. Material and methods

2.1. Animals and data

Animals (1,413 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

97 Blood samples were taken from the caudal vein (9 ml), kept below 4°C, and then
98 transported from the farm to the laboratory at Kasetsart University in Bangkok within 24
99 hours. The DNA from each sample was extracted and purified by applying a protocol of the
100 MasterPure^{TM} DNA Purification Kit (Epicentre®, USA). The quantity of DNA per sample
101 was measured using a NanoDrop 2000 (Thermo Fisher Scientific Inc., Wilmington, DE,
102 USA). The DNA was accepted as pure when the purity ratio is 260/280 of approximately 1.8,
103 and the DNA concentration was higher than 15 ng/µl.
104 The SNP genotyping was done by GeneSeek Inc. (Lincoln, NE, USA) using the
105 GeneSeek Genomic Profiler low density (GGP-LD) BeadChip that utilizes the Illumina
106 Infinium® chemistry (Illumina, San Diego, CA, USA). Each chip contains a total of 8,810
107 SNPs of which 8,305 SNP loci had known physical locations on the 29 autosomes (sex
108 chromosomes were ignored in this study). The SNPs with minor allele frequency (MAF) of less
109 than 0.05 were filtered out. After filtering, a total of 8,220 SNPs loci were included in the final
110 analysis.
111
112 2.3. Measures of linkage disequilibrium
113 Linkage disequilibrium (LD) is a measure of the non-random association between two
114 alleles that helps to infer the alleles at QTL that influence phenotypes of interest. Currently,
115 the most commonly used parameters to measure LD are D′ and r^2 (Zhao et al., 2005). The D′
116 is a measure of LD relative to the maximum possible value given the allele frequency of
117 SNPs. The D′ was considered from the frequencies of the haplotype of the SNP pairs, and it
118 was calculated as follows:

119 $$D' = \begin{cases} \dfrac{D}{min\,(f(A)\times f(b)\,,\, f(a)\times f(B))} & if\ D > 0 \\ \dfrac{D}{min\,(f(A)\times f(B)\,,\, f(a)\times f(b))} & if\ D < 0 \end{cases}$$

120 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 Roberson (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 10kb, 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) 1,059 cows (3/4 or 75%) and 7) 1,413 cows (1 or 100%). Cows for samples 1 to 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 with MAF < 0.05 and HWE (P < 0.0001) were excluded from the LD analyses. The LD values obtained from different sample sizes were compared. Values of D′ and r^2 from each sample size and regression coefficients of D′ and r^2 estimates on SNP pair distance (CORR procedure of SAS; SAS Inst., Inc., Cary, North Carolina, USA) were used to determine sample sizes that would provide reasonable LD estimates in the Thai multibreed dairy population.

3. Results and discussion

3.1. Descriptive summary of SNPs

A total of 8,220 (93%) SNPs met the filtering criteria (MAF $\geq$ 0.05). These markers covered 2,507.25 Mb of the genome; the shortest was chromosome 25 (42.85 Mb) and the longest was chromosome 1 (158.16 Mb). The density of SNPs varied across autosomes and ranged from 0.25 SNP/Mb (chromosome 20) to 0.34 SNP/Mb (chromosome 12). Furthermore, the distribution of SNPs across autosomes was not uniform, and they tended to be clustered in some regions of the chromosomes. Similar results were found in a Holstein population (Sargolzaei et al., 2008), Angus, Charolais and crossbred populations (Lu et al., 2012), and a Nellore population (Espigolan et al., 2013).

Almost 76% of the SNPs in the Thai population showed a MAF higher than 0.3 (Fig. 1). These SNPs tended to have a high MAF with a steep drop off towards rare alleles. The MAF distribution in this population was consistent with previous findings in *Bos taurus* cattle including Holstein (Sargolzaei et al., 2008; Kim and Kirkpatrick, 2009), Jersey, Brown Swiss (Wiggans et al., 2012), Fleckvieh, Dutch Black and White, Angus, Limousin and Charolais breeds (McKay et al., 2007; Pérez O'Brien et al., 2014). Conversely, a gradual decrease of MAF towards rare alleles has been observed in *Bos indicus* cattle including Nellore and

171 Brahman (Espigolan et al., 2013; McKay et al., 2007; Pérez O'Brien et al., 2014). This
172 indicated a substantial influence of *Bos taurus* genes in the Thai multibreed dairy population
173 (primarily high fractions of Holstein) resulting in a MAF distribution closer to that found in
174 *Bos taurus* than in *Bos indicus* breeds.

175 The average MAF in each chromosome ranged from 0.34 to 0.38, and the average
176 MAF across autosomes was 0.37 (Table 1). The average MAF found here were higher than
177 values reported for Holstein (0.28, Khatkar et al., 2008; 0.29, Bohmanova et al., 2010; 0.32,
178 Kim and Kirkpatrick, 2009; Wiggans et al., 2012), Jerseys (0.28, Wiggans et al., 2012) and
179 Brown Swiss cattle (0.29, Wiggans et al., 2012) and genetic diversity assessed from sequence
180 data by The Bovine HapMap Consortium (2009). The genetic variation within this Thai
181 multibreed dairy population may reflect the ancestral divergence among *Bos indicus* and *Bos*
182 *taurus* subspecies (McKay et al., 2007; Pérez O'Brien et al., 2014) as well as variation in
183 frequency and effect of alleles coming from the various component breeds.

184

185 3.2. The extent and decay of linkage disequilibrium

186 The extent of LD throughout the bovine genome plays an important role in
187 understanding the evolutionary biology (Mueller, 2004) and genome structure (Uimari et al.,
188 2005), and also its applications in gene mapping and genome-wide association studies
189 (Zapata, 2013; Raven et al., 2014). The level of LD between markers and QTL also affects
190 the accuracies of GEBV (Hayes et al., 2009). The mean LD between adjacent markers
191 averaged across all autosomes was 0.263 for D′ and 0.049 for r^2 (Table 1). These values were
192 slightly smaller than estimates elsewhere (Khatkar et al., 2008; Sargolzaei et al., 2008).
193 Variation in LD levels on autosomes is affected by recombination rate which in turn is
194 negatively associated with chromosome length (Farré et al., 2013). Thus, the LD levels in
195 longer chromosomes will extend for shorter distances, and consequently such chromosomes

196 have lower overall LD than shorter chromosomes (Bohmanova et al., 2010). The average LD
197 between adjacent SNPs on individual autosomes in this Thai population ranged from 0.207
198 (chromosome 28) to 0.368 (chromosome 20) for D′, and from 0.030 (chromosomes 27 and
199 22) to 0.090 (chromosome 16) for r^2 are show in Table 1. The average LD declined rapidly
200 with increasing physical distance between pairs of SNP to a very low level (Fig. 2). High LD
201 values were observed only at small distances between pairs of SNP.

202 Table 2 presents the frequency and mean for D′ and r^2 measured at different distances
203 between pairs of SNP up to a maximum of 5 Mb. The average of D′ for pairs of SNP located
204 at distances shorter than 10 kb was 0.904 and 85% of their pairs had D′ > 0.8. However, the
205 average D′ for pairs of SNPs located from 10 to 200 kb apart declined from 0.805 to 0.472,
206 and there was a decrease in D′ > 0.8 from 69% to 23%. The SNP pairs with D′ > 0.8 at short
207 distances decreased greatly when the distance between markers was more than 100 kb (Fig.
208 2). This indicated that not all markers separated distances smaller than 200 kb had D′ > 0.8
209 and that there was a gradual decline with increasing distance between markers. Values of D′
210 tended to decay more gradually than values of r^2 which had a clear exponential downward
211 trend with increasing physical distance (Fig. 2). The average r^2 was 0.515 and the proportion
212 of pairs that had $r^2 > 0.3$ was 61% for SNP pairs that were separated by 10 kb or less. Among
213 the SNP pairs 10 to 200 kb apart, the average r^2 declined rapidly from 0.321 to 0.116, and the
214 proportion of pairs that had $r^2 > 0.3$ decreased from 38% to 10%. At distances shorter than 60
215 kb, the proportion of markers with $r^2 > 0.3$ was 22%. This indicated that the r^2 values
216 declined rapidly with increasing distances between SNP in this population. Such decay of LD
217 was consistent with Khatkar et al. (2008), who reported that the average r^2 for pairs of SNPs
218 at small distances (< 40 kb) declined with increasing distances more rapidly than the average
219 D′.

The decay of LD in the genome with increasing physical distance showed extensive variability between genomic regions and chromosomes. This variation may be attributable to recombination rates that decreased as the length of chromosomes increased. Recombination rate are not uniformly distributed across each chromosome but clustered along chromosomal regions (Farré et al., 2013). Perhaps gene-conversion events contribute to this lack of uniformity (Ke et al., 2004).

Selection for traits of interest may affect the variability among genomic regions by increasing the frequency of certain alleles in the population. Thus, an increase of association between alleles at different loci would be linked to pairwise LD between high-frequency alleles (Pritchard and Przeworski, 2001; Ardlie et al., 2002). Therefore, useful LD ($r^2 > 0.3$) in this population would be those found at close distances (< 20 kb) in some regions of the genome. Different levels of LD have been indicated to be useful for different types of studies. Levels of r^2 above 0.3 increase the power to detect QTL in association studies (Ardlie et al., 2002).The SNP distances found in this study were longer than those found in indicine breeds (14 kb), and shorter than the taurine breeds (29 kb; Pérez O'Brien et al., 2014). This indicated that LD information from indicine and taurine populations may not be entirely applicable to this Thai *Bos indicus-Bos taurus* multibreed population.

Meuwissen et al. (2001) simulated the required level of LD (r^2) for genomic selection and achieved an accuracy of 0.85 for genomic breeding values for $r^2 = 0.2$. At this threshold ($r^2 = 0.2$), the distance between SNPs was less than 50kb in this Thai population, whereas it was 60 kb in a Holstein population in Australia (Khatkar et al., 2008) and North America (Bohmanova et al., 2010). Differences between dairy populations could be due to differences in genetic structure, sample sizes, measures of LD, marker types, marker densities and recent history of the population (Pritchard and Przeworski, 2001). Each of these factors could affect

244 the estimation of LD. In particular, D′ values could be overestimated at low allele frequencies
245 and in small sample sizes (Bohmanova et al., 2010; Espigolan et al., 2013).
246
247 3.3. Effect of MAF and sample size on the extent of linkage disequilibrium
248 Four different minimum MAF thresholds (0.05, 0.10, 0.15 and 0.20) were used to
249 assess the effect of allele frequencies on LD (D′ and r^2). The average D′ for pairs of SNPs at a
250 distance of more than 10 kb apart decreased when the MAF threshold increased (Fig. 3). In
251 contrast, the average r^2 increased when the MAF threshold was increased, particularly at
252 short distances (0 to 50 kb; Fig. 4). For SNP pairs closer than 10 kb, the average D′ was 0.904
253 when MAF > 0.05, and decreased to 0.898 when MAF increased or was larger than 0.20.
254 This was different from the average r^2, which was 0.515 when MAF > 0.05 and it was higher
255 (0.599) when MAF > 0.20. These results were similar to those from previous studies where
256 D′ overestimated the extent of LD especially in cases of low MAF values (Bohmanova et al.,
257 2010; Espigolan et al., 2013). This may be due to the value of the denominator in the formula
258 of D′ which is equal to the minimal product of SNP allele frequencies (Bohmanova et al.,
259 2010). Thus, it is likely that SNP pairs with low allele frequencies yielded inflated of D′
260 values in this Thai population, whereas the opposite occurred for pairs with high allele
261 frequencies.
262 The effect of sample size on accuracy in estimation of D′ is shown in Fig. 5. With
263 small sample sizes, the D′ estimates tended to deviate from the estimates of the complete
264 dataset (1,413 cows). Differences were more noticeable for LD measured between SNP
265 markers at distances greater than 60 kb. Conversely, r^2 estimates were only slightly affected
266 by a decrease in sample size (Fig. 6). This indicated that estimates of D′ were more dependent
267 on sample size than estimates of r^2. Estimates of D′ from samples larger than 177 animals had
268 minimal deviation from D′ in the complete population. Thus, sample sizes of 177 and above

would need to be used to estimate D′ in this Thai population. On the other hand, r^2 values from sample sizes larger than 89 differed only slightly from the r^2 value in the complete dataset, indicating that r^2 was barely influenced by sample size. The only sample size that resulted in an overestimate of r^2 was 45 animals.

Correlations between accuracies of estimation of the extent of LD (D′ and r^2) obtained in sample sizes 1 to 6 and the complete dataset are shown in Table 3. Correlation estimates for D′ higher than 0.9 needed a minimum sample size of 177 cows. However, estimates of r^2 with accuracies larger than 0.9 required a minimum sample size of only 89 cows. The D′ estimates show much more inflation in small samples than r^2 estimates, which was similar to Bohmanova et al. (2010), who reported that minimum sample sizes were 444 bulls to estimate D′ and 55 bulls to estimate r^2 in North American Holstein. Similarly, Khatkar et al. (2008) indicated 400 bulls were necessary to estimate D′ and 75 bulls were required to estimate r^2 in Australian Holstein-Friesian.

Values of r^2 may be more useful than D′ to estimate LD in terms of the power to detect associations in genome-wide association studies because sample size is usually a limiting factor of these studies and increasing sample size to compensate for weak LD may be impractical (Ardlie et al., 2002). Further, r^2 is a more robust measure of LD than D′ because it is less sensitive to allele frequencies and to small sample sizes (Bohmanova et al., 2010).

4. Conclusions

The level of LD estimated for pairs of SNPs at short distances (< 50 kb) showed higher LD than pairs of SNPs at greater distances (> 50 kb) in 1,413 Thai multibreed dairy cattle genotyped for 8,220 SNPs. The D′ measure of LD was strongly inflated in small samples and at low allele frequencies. Hence, r^2 is should be the measure of choice for fine

mapping, identification of haplotype blocks, and finding the correct physical location of selection markers. Results from this research will be useful for genome-wide association studies and genomic selection of dairy cattle in tropical regions.

Acknowledgments

This manuscript is part of research projects co-funded by the National Science and Technology Development Agency (NSTDA), Kasetsart University, and the Dairy Farming Promotion Organization. The authors thank the NSTDA, University, and Industry Research Collaboration (NUIRC) for giving a scholarship to the first author. We also thank Thai dairy farmers, dairy cooperatives, and private organizations for their participation and support. The authors declare that they do not have any conflict of interests.

References

Ardlie, K.G., Kruglyak, L., Seielstad, M., 2002. Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. 3, 299-309.

Barrett, J.C., Fry, B., Maller, J., Daly, M.J., 2005. Haploview analysis and visualization of LD and haplotype maps. Bioinformatics 21 263-265.

Bohmanova, J., Sargolzaei, M., Schenkel, F.S., 2010. Characteristics of linkage disequilibrium in North American Holsteins. BMC Genomics 11, 421.

Espigolan, R., Baldi, F., Boligon, A.A., Souza, F.R.P., Gordo, D.G.M., Tonussi, R.L., Cardoso, D.F., Oliveira, H.N., Tonhati, H., Sargolzaei, M., Schenkel, F.S., Carvalheiro, R., Ferro, J.A., Albuquerque, L.G., 2013. Study of whole genome linkage disequilibrium in Nellore cattle. BMC Genomics 14, 305.

317 Farré, M., Michelletti, D., Ruiz-Herrera A., 2013. Recombination rates and genomic shuffling
318 in Human and Chimpanzee – a new twist in the chromosomal speciation theory. Mol.
319 Biol. Evol. 30, 853-864.
320 Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Goddard, M.E., 2009. Invited review:
321 Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92, 433-443.
322 Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor. Appl.
323 Genet. 38, 226-231.
324 Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J., Whittaker, P., Collins,
325 A., Morris, A.P., Bentley, D., Cardon, L.R., Deloukas, P., 2004. The impact of SNP
326 density on fine-scale patterns of linkage disequilibrium. Hum. Mol. Genet. 13, 577-
327 588.
328 Khatkar, M.S., Nicholas, F.W., Collins, A.R., Zenger, K.R., Cavanagh, J.A.L., Barris, W.,
329 Schnabel, R.D., Taylor, J.F., Raadsma, H.W., 2008. Extent of genome-wide linkage
330 disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP
331 panel. BMC Genomics. 9, 187.
332 Kim, E.S., Kirkpatrick, B.W., 2009. Linkage disequilibrium in the North American Holstein
333 population. Anim. Genet. 40, 279-288.
334 Koonawootrittriron, S., Elzo, M.A., Thongprapi, T., 2009. Genetic trends in a Holstein x
335 other breeds multibreed dairy population in Central Thailand. Livest. Sci. 122, 186-
336 192.
337 Lewontin, R.C., 1964. The interaction of selection and linkage. I. General considerations;
338 heterotic models. Genet. 49, 49-67.
339 Lu, D., Sargolzaei, M., Kelly, M., Li, C., Voort, G.V., Wang, Z., Plastow, G., Moore, S.,
340 Miller, S.P., 2012. Linkage disequilibrium in Angus, Charolais and Crossbred beef
341 cattle. Front. Genet. 3, 152.

342 McKay, S.D., Schnabel, R.D., Murdoch, B.M., Matukumalli, L.K., Aerts, J., Coppieters, W.,
343 Crews, D., Neto, E.D., Gill, C.A., Gao, C., Mannen, H., Stothard, P., Wang, Z., Van
344 Tassell, C.P., Williams, J.L., Taylor, J.F., Moore, S.S., 2007. Whole genome linkage
345 disequilibrium maps in cattle. BMC Genet. 8, 74.
346 Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using
347 genome-wide dense marker maps. Genet. 157, 1819-1829.
348 Mueller, J.C., 2004. Linkage disequilibrium for different scales and applications. Brief.
349 Bioinform. 5, 355-364.
350 Pérez O'Brien, A.M., Meszaros, G., Utsunomiya, Y.T., Sonstegard, T.S., Garcia, J.F.,
351 VanTassell, C.P., Carvalheiro, R., da Silva, M.V.B., Solkner, J., 2014. Linkage
352 disequilibrium levels in *Bos indicus* and *Bos taurus* cattle using medium and high
353 density SNP chip data and different minor allele frequency distributions. Livest. Sci.
354 166, 121-132.
355 Pritchard, J.K., Przeworski, M., 2001. Linkage disequilibrium in Human: Model and data.
356 Amer. J. Hum. Genet. 69, 1-14.
357 Qanbari, S., Pimental, E.C.G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R., Simianer,
358 H., 2010. The pattern of linkage disequilibrium in German Holstein cattle. Anim.
359 Genet. 41, 346-356.
360 Raven, L.A., Cocks, B.G., Hayes, B.J., 2014. Multibreed genome wide association can
361 improve precision of mapping causation variants underlying milk production in dairy
362 cattle. BMC Genomics 15, 62.
363 Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.L., Lavery, T.,
364 Kouyoumjian, R., Farhadian, S.F., Ward, R., Lander, E.S., 2001. Linkage
365 disequilibrium in the human genome. Nature 411, 199-204.

366 Sargolzaei, M., Schenkel, F.S., Jansen, G.B., Schaeffer, L.R., 2008. Extent of linkage
367 disequilibrium in Holstein cattle in North American. J. Dairy Sci. 91, 2106-2117.
368 Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Meuwissen, T.H.E., 2008 Genomic selection
369 using different marker types and densities. J. Anim. Sci. 86, 2447-2454.
370 Teare, M.D., Dunning, A.M., Durocher, F., Rennart, G., Easton, D. F., 2002. Sampling
371 distribution of summary linkage disequilibrium measures. Ann. Hum. Genet. 66, 223-
372 233.
373 The Bovine HapMap Consortium., 2009. Genome-wide survey of SNP variation uncovers the
374 genetic structure of cattle breeds. Science 324, 528-532.
375 Uimari, P., Kontkanen, O., Visscher, P. M., Pirskanen, M., Fuentes, R., Salonen, J. T., 2005.
376 Genome-wide linkage disequilibrium from 100,000 SNPs in the East Finland Founder
377 Population. Twin Res. Hum. Genet. 8, 185-189.
378 Varoneze, R., Bastiaansen, J.M.W., Knol, E.F., Guimaraes, S.E.F., Silva, F.F., Harlizius, B.,
379 Lopes, M.S., Lopes, P.S., 2014. Linkage disequilibrium patterns and persistence of
380 phase in purebred and crossbred pig (*Sus scrofa*) populations. BMC Genet. 15, 126.
381 Wiggans, G.R., Cooper, T.A., Van Raden, P.M., Olson, K.M., Tooker, M.E., 2012. Use of
382 the Illumina Bovine3K BeadChip in dairy genomic evaluation. J. Dairy Sci. 95, 1552-
383 1558.
384 Yang, J., Zhu, W., Chen, J., Zheng, Q., Wu, S., 2014. Genome-wide two-marker linkage
385 disequilibrium mapping of quantitative trait loci. BMC Genet. 15, 20.
386 Zapata, C., 2013. Linkage disequilibrium measures for fine-scale mapping of disease loci are
387 revisited. Front. Genet. 4, 228.
388 Zhao, H., Nettleton, D., Soller, M., Dekkers, J.C., 2005. Evaluation of linkage disequilibrium
389 measures between multi-allelic markers as predictors of linkage disequilibrium
390 between markers and QTL. Genet. Res. 86, 77-87.

391 **Table 1** Descriptive summary of SNPs obtained for each autosome in the Thai multibreed dairy population

Chrom	Chrom Length (Mb)	SNPs (n)	Distance mean ± SD (Mb)	Median D′	Mean D′ ± SD	Median r^2	Mean r^2 ± SD	MAF mean ± SD
1	158.162	530	2.350 ± 1.496	0.206	0.278 ± 0.246	0.020	0.056 ± 0.112	0.370 ± 0.093
2	136.484	433	2.436 ± 1.458	0.218	0.270 ± 0.220	0.020	0.046 ± 0.078	0.367 ± 0.104
3	121.375	405	2.219 ± 1.550	0.226	0.297 ± 0.257	0.022	0.065 ± 0.135	0.367 ± 0.099
4	120.615	359	2.532 ± 1.413	0.189	0.228 ± 0.179	0.016	0.035 ± 0.050	0.375 ± 0.093
5	120.784	380	2.339 ± 1.506	0.212	0.305 ± 0.280	0.019	0.051 ± 0.107	0.364 ± 0.111
6	121.357	444	2.424 ± 1.432	0.193	0.247 ± 0.206	0.017	0.042 ± 0.082	0.370 ± 0.103
7	112.610	347	2.480 ± 1.427	0.185	0.236 ± 0.194	0.016	0.040 ± 0.072	0.371 ± 0.098
8	113.321	347	2.521 ± 1.440	0.199	0.243 ± 0.196	0.019	0.043 ± 0.070	0.376 ± 0.092
9	105.463	326	2.393 ± 1.477	0.216	0.281 ± 0.239	0.021	0.048 ± 0.089	0.363 ± 0.105
10	104.215	320	2.492 ± 1.420	0.201	0.244 ± 0.191	0.018	0.043 ± 0.065	0.371 ± 0.098
11	107.043	340	2.479 ± 1.421	0.188	0.230 ± 0.185	0.016	0.035 ± 0.059	0.368 ± 0.098
12	91.092	267	2.444 ± 1.438	0.180	0.232 ± 0.197	0.015	0.037 ± 0.071	0.371 ± 0.100
13	84.149	270	2.434 ± 1.439	0.203	0.254 ± 0.208	0.017	0.042 ± 0.070	0.362 ± 0.104
14	84.616	275	2.486 ± 1.422	0.225	0.265 ± 0.205	0.022	0.047 ± 0.074	0.371 ± 0.099
15	85.012	270	2.428 ± 1.462	0.170	0.222 ± 0.198	0.014	0.039 ± 0.083	0.381 ± 0.092
16	80.925	298	1.990 ± 1.585	0.261	0.359 ± 0.305	0.026	0.090 ± 0.178	0.348 ± 0.111
17	74.966	222	2.484 ± 1.419	0.189	0.234 ± 0.189	0.016	0.038 ± 0.066	0.370 ± 0.103
18	65.979	231	2.094 ± 1.547	0.175	0.242 ± 0.229	0.014	0.036 ± 0.075	0.368 ± 0.096
19	64.007	217	2.396 ± 1.476	0.166	0.224 ± 0.211	0.012	0.044 ± 0.112	0.371 ± 0.098
20	71.794	284	2.062 ± 1.625	0.274	0.368 ± 0.305	0.026	0.088 ± 0.180	0.344 ± 0.119
21	70.608	248	2.138 ± 1.566	0.226	0.307 ± 0.268	0.021	0.061 ± 0.127	0.357 ± 0.102
22	60.931	199	2.449 ± 1.430	0.171	0.211 ± 0.175	0.013	0.030 ± 0.051	0.375 ± 0.099
23	52.129	203	2.175 ± 1.562	0.191	0.268 ± 0.252	0.017	0.067 ± 0.152	0.375 ± 0.095
24	62.644	212	2.493 ± 1.423	0.180	0.224 ± 0.182	0.015	0.036 ± 0.065	0.372 ± 0.097
25	42.851	160	2.367 ± 1.451	0.179	0.238 ± 0.213	0.014	0.035 ± 0.061	0.365 ± 0.097
26	51.680	173	2.417 ± 1.457	0.180	0.230 ± 0.198	0.013	0.031 ± 0.048	0.355 ± 0.108
27	45.369	153	2.437 ± 1.441	0.169	0.221 ± 0.196	0.013	0.030 ± 0.053	0.372 ± 0.104
28	46.102	148	2.413 ± 1.434	0.161	0.207 ± 0.179	0.013	0.031 ± 0.061	0.374 ± 0.102
29	50.972	159	2.443 ± 1.419	0.175	0.217 ± 0.183	0.015	0.036 ± 0.069	0.375 ± 0.087
Overall	2,507.255	8,220	2.360 ± 1.486	0.200	0.263 ± 0.231	0.018	0.049 ± 0.018	0.368 ± 0.101

392 **Table 2** Frequency and mean LD (D′ and r^2) between SNPs at different distances pooled over all autosomes

Distance	SNP pair (n)	Median D′	Mean D′ ± SD	Median r^2	Mean r^2 ± SD	D′ > 0.8	r^2 > 0.3
0 to 10 kb	997	1.000	0.904 ± 0.208	0.514	0.515 ± 0.361	844 (84.65)*	608 (60.98)**
10 to 20 kb	841	0.972	0.805 ± 0.286	0.189	0.321 ± 0.313	583 (69.32)	323 (38.41)
20 to 30 kb	989	0.923	0.768 ± 0.296	0.172	0.278 ± 0.272	631 (63.80)	330 (33.37)
30 to 40 kb	834	0.894	0.730 ± 0.308	0.147	0.246 ± 0.265	476 (57.07)	244 (29.26)
40 to 50 kb	749	0.809	0.694 ± 0.315	0.109	0.202 ± 0.226	377 (50.33)	178 (23.77)
50 to 60 kb	699	0.775	0.644 ± 0.353	0.097	0.188 ± 0.229	336 (48.07)	157 (22.46)
60 to 70 kb	642	0.591	0.572 ± 0.352	0.089	0.167 ± 0.213	240 (37.38)	114 (17.76)
70 to 80 kb	568	0.538	0.548 ± 0.337	0.086	0.164 ± 0.209	169 (29.75)	109 (19.19)
80 to 90 kb	510	0.522	0.529 ± 0.355	0.061	0.150 ± 0.211	161 (31.57)	85 (16.67)
90 to 100 kb	502	0.464	0.511 ± 0.326	0.065	0.127 ± 0.172	132 (26.29)	57 (11.35)
100 to 200 kb	3,259	0.407	0.472 ± 0.324	0.053	0.116 ± 0.166	763 (23.41)	329 (10.10)
200 to 300 kb	2,257	0.298	0.351 ± 0.265	0.036	0.070 ± 0.092	184 (8.15)	71 (3.15)
300 to 400 kb	2,455	0.264	0.315 ± 0.241	0.030	0.062 ± 0.084	125 (5.09)	51 (2.08)
400 to 500 kb	2,632	0.260	0.306 ± 0.232	0.028	0.057 ± 0.077	109 (4.14)	53 (2.01)
500 to 600 kb	2,505	0.246	0.293 ± 0.224	0.026	0.053 ± 0.069	93 (3.71)	30 (1.20)
600 to 700 kb	2,689	0.230	0.274 ± 0.209	0.023	0.046 ± 0.061	55 (2.05)	25 (0.93)
700 to 800 kb	2,720	0.227	0.278 ± 0.214	0.023	0.047 ± 0.062	75 (2.76)	21 (0.77)
800 to 900 kb	2,640	0.224	0.266 ± 0.202	0.021	0.046 ± 0.061	50 (1.89)	21 (0.80)
900 to 1,000 kb	2,673	0.217	0.260 ± 0.203	0.021	0.043 ± 0.057	50 (1.87)	14 (0.52)
1 to 2 Mb	25,534	0.207	0.247 ± 0.189	0.019	0.039 ± 0.052	317 (1.24)	87 (0.34)
2 to 3 Mb	24,956	0.185	0.224 ± 0.176	0.015	0.032 ± 0.043	188 (0.75)	36 (0.14)
3 to 4 Mb	24,420	0.169	0.208 ± 0.166	0.013	0.027 ± 0.037	150 (0.61)	6 (0.02)
4 to 5 Mb	23,846	0.155	0.191 ± 0.157	0.011	0.023 ± 0.031	115 (0.48)	6 (0.03)

393 * Percentage of pairs of SNPs with D′ > 0.8; ** Percentage of pairs of SNPs with r^2 > 0.3

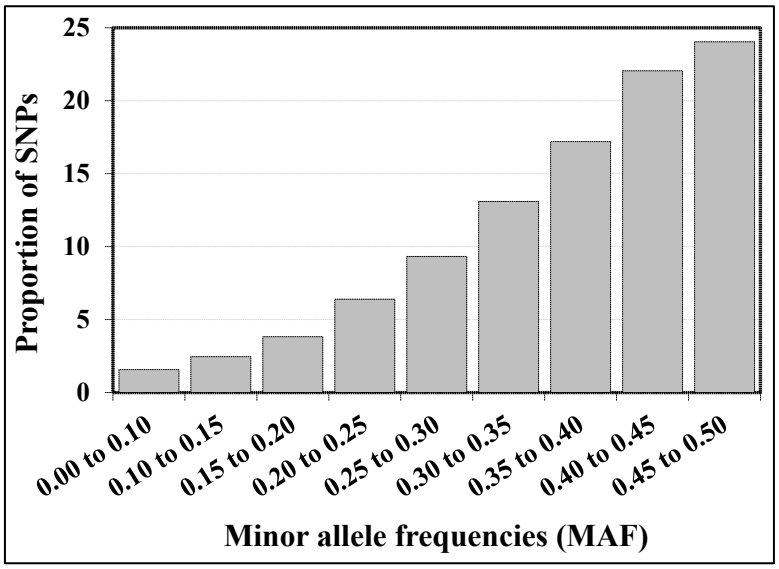
394 **Table 3** Correlations between D′ and r^2 estimates from six sample sizes and D′ and r^2

395 estimates from the complete dataset (1,413 cows)

Sample Size (cows)	D′	r^2
1,059	0.994	0.999
707	0.984	0.997
354	0.957	0.990
177	0.901	0.973
89	0.810	0.941
45	0.688	0.882

396 [1]All correlations were significant ($P < 0.0001$)

397



398

399 **Fig. 1** Distribution of proportion of SNPs by minor allele frequency (MAF) after quality

400 control

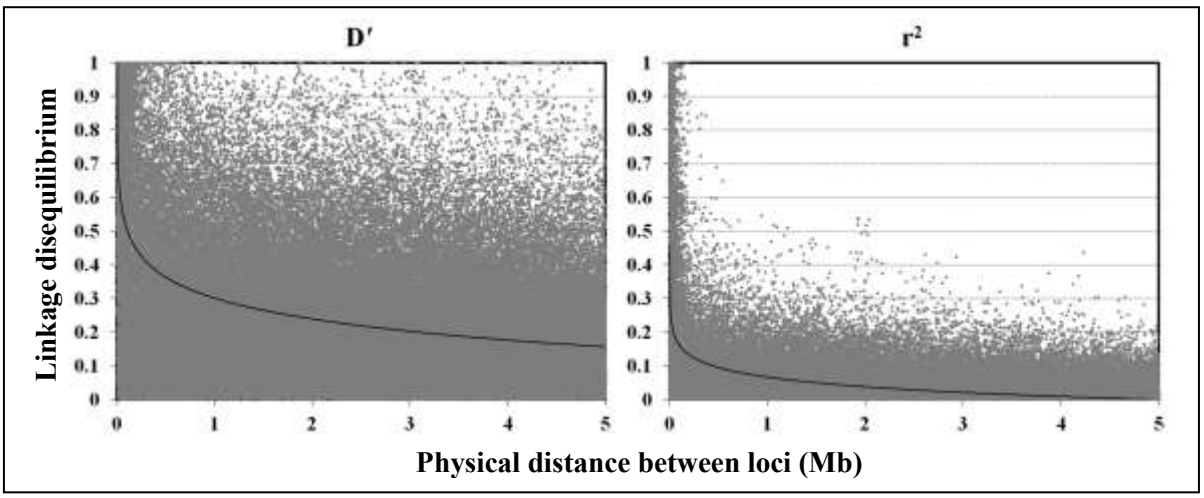


Fig. 2 Distribution of linkage disequilibrium measurements (D′ and r^2) in relation to physical distance among loci (Mb) across autosomes

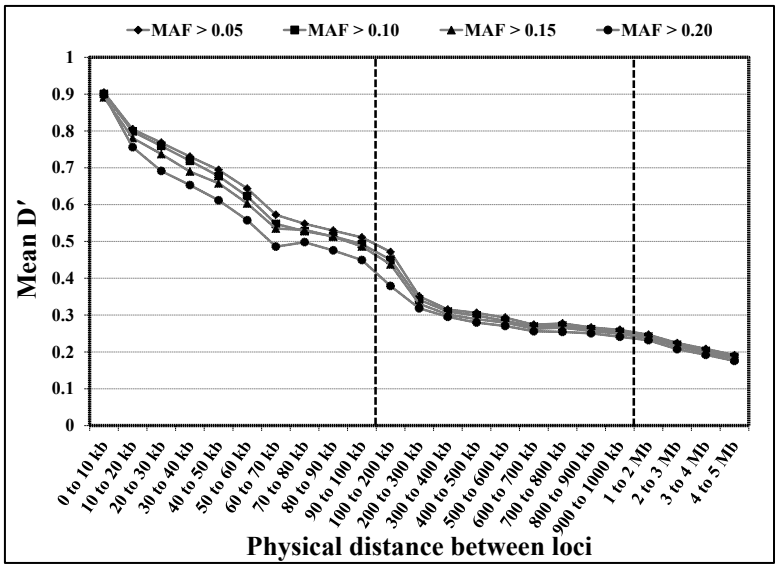


Fig. 3 Mean D′ at different physical distances pooled across autosomes for different thresholds of minor allele frequency (MAF)

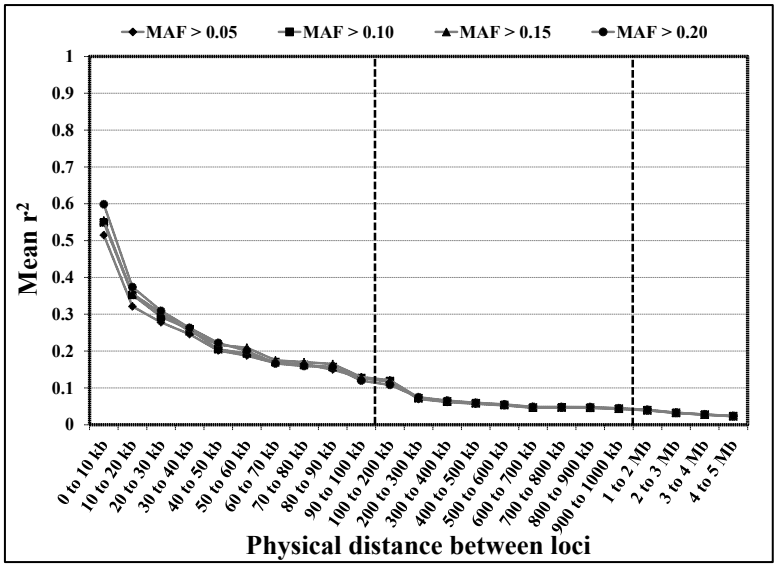


Fig. 4 Mean r^2 at different physical distances pooled across autosomes for different thresholds of minor allele frequency (MAF)

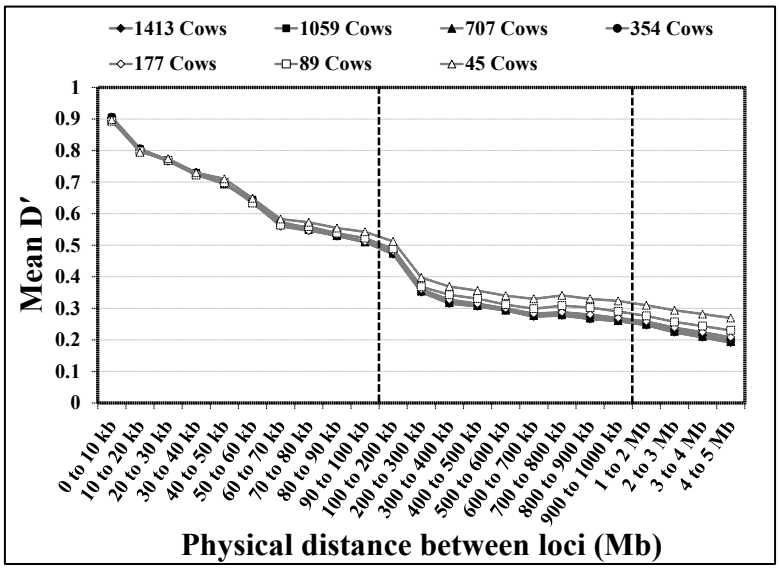
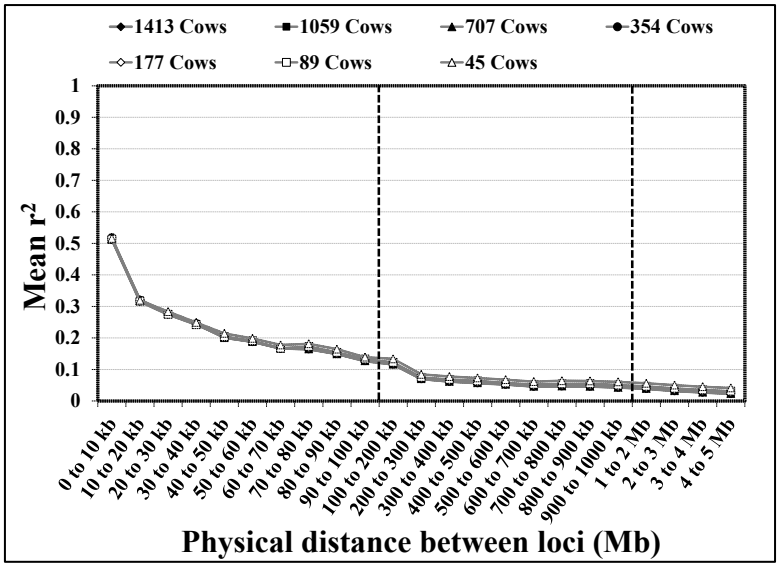


Fig. 5 Distribution of the mean D′ at different physical distances for different sample sizes



425

426

427 **Fig. 6** Distribution of the mean r^2 at different the physical distances for different sample sizes