Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows

Citation for published version:
Banos, G & Coffey, MP 2010, 'Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows' Journal of Dairy Science, vol 93, no. 6, pp. 2775-2778., 10.3168/jds.2009-2613

Digital Object Identifier (DOI):
10.3168/jds.2009-2613

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher final version (usually the publisher pdf)

Published In:
Journal of Dairy Science

J. Dairy Sci. 93:2775–2778
doi:10.3168/jds.2009-2613

Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows

G. Banos*[1] and M. P. Coffey†
*Department of Animal Production, Faculty of Veterinary Medicine, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece
†Sustainable Livestock Systems, Scottish Agricultural College, Bush Estate, Midlothian, EH26 0PH, United Kingdom

ABSTRACT

The objective of this study was to describe results of a genome-wide map of single nucleotide polymorphisms (SNP) and assess the linkage disequilibrium (LD) level in 2 divergent selection lines of dairy cows. DNA extracted from 299 Holstein cows was used to determine genotypes in 54,001 SNP loci using the BovineSNP50 array (Illumina Inc., San Diego, CA). Animals were from 2 genetic lines (166 genetically selected for fat and protein yield vs. 133 controls) raised on an experimental farm. Data edits removed loci with a major allele frequency greater than 0.95, genotypes in fewer than 100 cows, and missing valid chromosomal assignment or position. After edits, 41,859 loci (77.5% of the original total) were kept for further analysis. Linkage disequilibrium (LD) values were calculated for all possible syntenic SNP locus pairs located within intervals of 1 million base pairs, as the squared correlation between alleles. Pairwise haplotypes were determined using parsimony. Linkage disequilibrium was calculated for all animals and then for each genetic line separately. The average LD calculated across all chromosomes was 0.069, 0.071, and 0.075 for all, control, and select line cows, respectively. Genetic line had a statistically significant effect on LD. Of all locus pairs studied, 53,487 to 95,279 (depending on the data set) were in LD >0.30, which may be considered the minimum useful for mapping purposes and genomic selection. Useful LD was mostly found between adjacent pairs located within 30,000 to 50,000 bases. A few locus pairs (844–1,070 in the 3 data sets) were found in almost perfect (>0.99) LD. The overall product-moment correlation of LD values between the control and select lines was 0.79 (significantly different from 1), ranging from 0.71 to 0.84 for different chromosomes. Looking at this correlation by SNP pair distance revealed that persistence of LD phase across the 2 lines extended chiefly for 200,000 bases. Selection is likely to have affected LD in the studied cow population. These results may be useful to gene detection and genome-wide association studies.

Key words: linkage disequilibrium, genome-wide scan, selection lines

The advent of high-throughput genotyping technology has enabled large-scale determination of specific marker locus genotypes throughout an animal's genome. Currently, a bovine DNA array comprising approximately 50,000 SNP distributed across the genome is commercially available. This development paves the way for genome-wide association studies, thereby facilitating gene detection and mapping as well as genomic evaluation and selection for complex traits (Goddard and Hayes, 2009). All such applications require an assessment of the linkage disequilibrium (**LD**) state between genetic markers.

The objective of the present study was to describe results from a genome-wide map and examine the LD in 2 divergent selection lines of dairy cattle.

Animals considered were Holstein cows raised on a research farm of the Scottish Agricultural College at Crichton, Dumfries, Scotland (55°02′N, 3°34′W, 40 m altitude). The herd was established in 1973. In the original herd, cows had been randomly selected from the best available on the market at the time and assigned to a select line; these represented daughters of the top 10 bulls for fat and protein yield. United Kingdom national average cows were also purchased and assigned to a control line. Select line animals were mated to bulls of the highest EBV for fat and protein yield, where these bulls had only previously been used in the herd for 1 yr and inbreeding coefficients in resulting calves would be acceptable (<6.25%). Selected sires were generally within the top 10 bulls available nationally. Control line cows were mated to bulls whose estimated genetic merit for fat and protein yield was the UK national average. Cows participating in the present study were a subsample of 299 animals born in the period 1993 to 2005 and represented cows that had been subject to varying generations of selection. In all, 133 cows belonged to the control line and 166 to the select line.

Received August 3, 2009.
Accepted March 4, 2010.
[1]Corresponding author: banos@vet.auth.gr

DNA was extracted from these cows and used to determine genotypes in 54,001 SNP loci distributed across the entire genome with the Illumina BovineSNP50 array (Illumina Inc., San Diego, CA). Data edits removed loci missing valid chromosomal assignment or position and having legitimate genotypes in fewer than 100 cows. An additional edit removed all loci with a major allele frequency >0.95. The Hardy-Weinberg equilibrium (**HWE**) state was assessed in all remaining SNP loci using a chi-squared test. Observed heterozygosity was compared with the theoretical expectation based on allelic frequency.

Pairwise haplotypes, needed to calculate LD, were constructed using parsimony, which entails determination and testing of haplotype sets using genotype observation and custom software. Parsimony is based on the assignment of haplotypes to genotypes that minimizes the number of unresolved haplotypes (Clark, 1990). In practice, definite and unambiguous haplotypes are identified first from homozygous animals and then from individuals with only one heterozygous SNP locus (Tier, 2006). Following that, unique pairs of haplotypes are assigned, where possible, that are compatible with their genotypes. In the present study, an algorithm was developed to test haplotype pairs against genotypes observed. Putative haplotypes were confirmed with custom software. In addition, 72 pairs of genotyped dams and daughters were used to verify assigned haplotypes. Remaining ambiguous cases (<8%) were removed from further analysis.

Linkage disequilibrium among all syntenic SNP locus pairs situated within 1 million base (Mb) intervals, corresponding roughly to 1 cM, was calculated as the squared correlation between alleles at 2 loci (Hill and Robertson, 1968). The latter is considered the most suitable measure for calculating LD between biallelic markers such as SNP (Zhao et al., 2007). Sliding windows of 1 Mb were visualized alongside each chromosome. Linkage disequilibrium was calculated among all SNP pairs within a 1-Mb window, which was then slid ahead by one SNP and LD was calculated again among the new pairs. The SNP considered at this stage were those remaining after all edits described previously that were also in HWE. When the process yielded repeat estimates of LD between 2 locus pairs, their average was used as the final LD estimate. This exercise was repeated for all 30 bovine chromosome pairs. In addition, LD values were calculated within genetic line (for select and control cows separately).

Marker loci were distributed throughout the genome, ranging from a minimum of 747 SNP on chromosome X to a maximum of 3,343 SNP on chromosome 1. On average, there were 17.5 SNP per Mb with the highest concentrations on chromosomes 1, 2, and 4. About 3% of the loci were removed from further examination because of missing chromosomal position or having genotypes in fewer than 100 cows. Furthermore, 10,394 loci were removed because they had a major allele frequency >0.95. The latter were considered practically monomorphic and would require very large sample sizes for detection of small differences.

After the above edits, 41,859 SNP loci (77.5% of the original total) were kept for further analysis. The majority (96%) were found in HWE ($P > 0.05$). Compared with theoretical values, average heterozygote deficit was −0.005 (SD 0.023, range −0.105 – 0.081) and 0.012 (SD 0.137, range −0.497 – 0.460) for loci that were and were not in HWE, respectively.

Average LD, calculated between all SNP locus pairs found in HWE and located within 1-Mb intervals, was 0.069 (SD 0.108), 0.071 (SD 0.111), and 0.075 (SD 0.118) for all, control and select line cows, respectively. These results are mean LD calculated in all 1-Mb intervals within a chromosome and averaged across all chromosomes. The LD values obtained here are somewhat lower than estimates from another high-density map study of Holstein cattle (Sargolzaei et al., 2008). The latter considered a North American Holstein proven bull population (497 animals), which may be more intensely selected than the cow resource population used in the present study.

Figure 1 depicts LD results by genetic line and chromosome number. In all cases, average LD was higher in the select line than in the control line.

The effect of genetic line on average LD was assessed by using a linear model that fitted the effects of line and chromosome number. The R^2 value for this model was 0.96 and both effects were highly statistically significant ($P < 0.001$). The marginal difference between select and control lines, adjusted for the effect of chromosome number, was 0.004 ± 0.0001. This may be attributed to initial formation of and ongoing sire selection for production traits in the select line. Linkage disequilibrium is expected to be stronger and extend over larger distances in genetically selected animals compared with moderately selected or unselected populations (Thévenon et al., 2007). Of course, genetic drift because of finite population size may also contribute to LD differences. However, the randomness of genetic drift could shift this difference in favor of either the select or the control line. In the present study, LD was consistently higher in the select line (Figure 1).

Of all SNP locus pairs examined, 53,487, 73,708, and 95,279 in the entire data set and the control and select line subsets, respectively, were in LD >0.30, which may be considered the minimum useful for mapping purposes and genomic selection (Meuwissen et al., 2001; Sargolzaei et al., 2008). Although somewhat arbitrary,

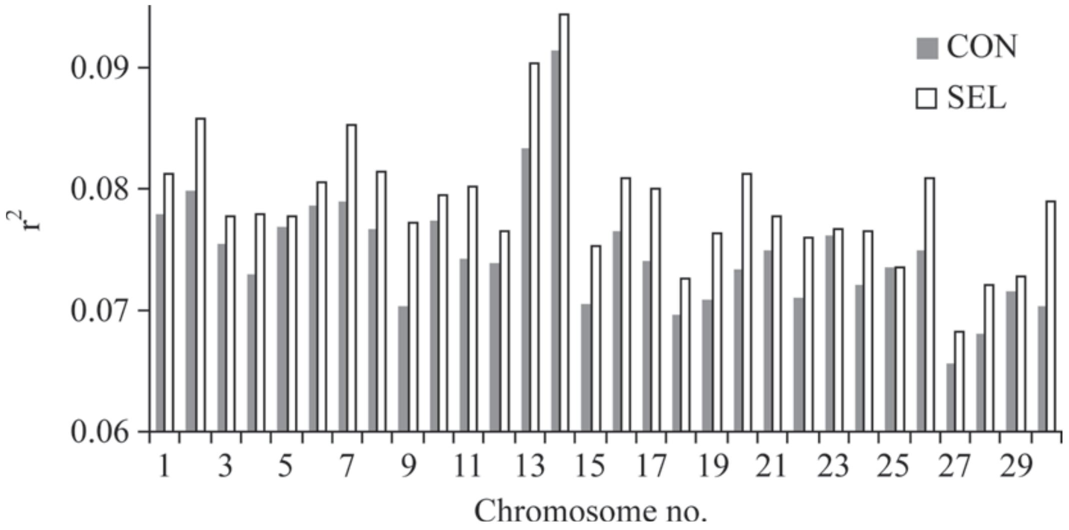

Figure 1. Average linkage disequilibrium (r^2) between SNP loci situated within 1 million base (1 Mb) intervals, by chromosome number; CON = control, SEL = select line of cows.

this threshold value is associated with an accuracy of genomic prediction >0.85. Furthermore, in terms of statistical power in genetic association studies, the sample size must increase by 1/r^2 for using a SNP to achieve the same power as using an actual gene (Pritchard and Przeworski, 2001); in this regard, an r^2 value of 0.30 assumes a maximum 3-fold increase in sample size in Holstein cattle (Sargolzaei et al., 2008).

A few SNP locus pairs (844, 984, and 1,070 in all, control, and select cows, respectively) were found in almost perfect (>0.99) LD, implying that one locus in the pair would be sufficient to capture a functional gene.

Looking at adjacent SNP loci, average distance was 0.057 Mb and LD was 0.125 (SD 0.216) in the entire data set; slightly higher LD values were observed within the control and select lines. Figure 2 illustrates average LD between such pairs in relation to their distance, considering all cows in the data set; as expected, LD decreased as distance increased. Similar LD decay was observed in the control and select lines. Meaningful LD (>0.30) was found between loci situated within 30,000 to 50,000 bases. In this respect, genotyping with a 50K SNP array appears to provide, on average, marginal coverage of the genome; denser marker genotyping would result in more loci being situated within the desirable range.

The overall product-moment correlation of LD calculated between the control and select lines of cows was 0.79 (statistically different from 1; $P < 0.05$). This correlation refers to the correlation of LD values of SNP pairs in one line with LD values of the same pairs in

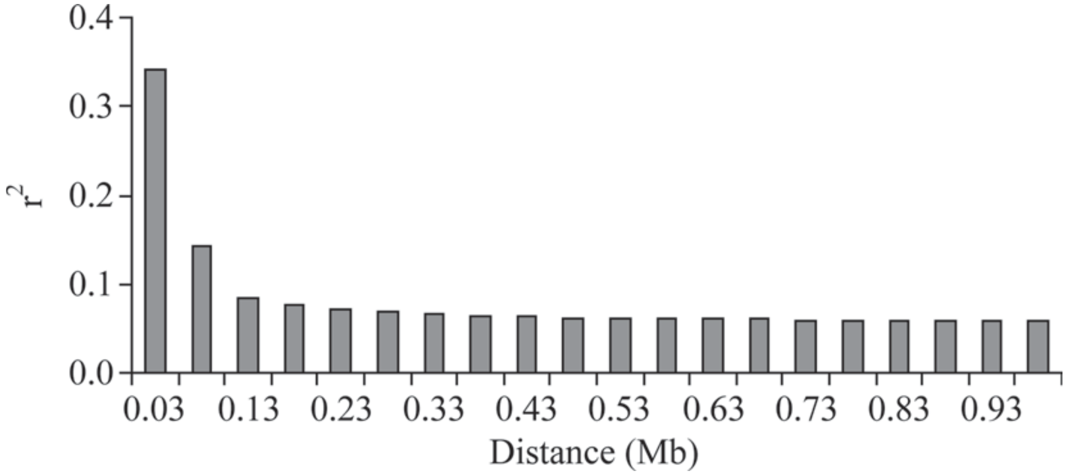

Figure 2. Average linkage disequilibrium (r^2) between SNP loci in relation to their distance (million bases, Mb); cows from both lines were considered.

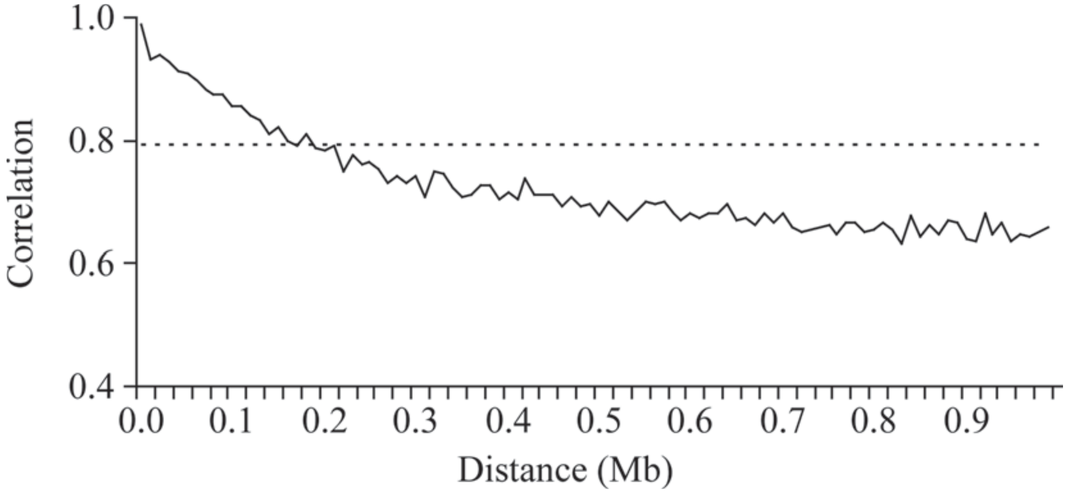

Figure 3. Correlation of linkage disequilibrium (r^2) calculated between the control and select lines of cows by distance (million bases, Mb) between SNP pairs (solid line); dotted line represents overall LD correlation between the 2 lines.

the other line. The exercise considered all SNP pairs located within 1-Mb intervals. By chromosome number, estimates ranged from 0.71 (chromosome X) to 0.84 (chromosome 14). To better assess the utility of this result, the control line was randomly divided into 2 halves and the LD correlation between the 2 subgroups was calculated. This exercise was repeated 100 times. The average LD correlation of the 100 replicates between subgroups of the control line was 0.94 (SD 0.07). These results support the notion that divergent selection, practiced since the initial formation of the studied population, has influenced LD in different ways in the 2 genetic lines.

Persistence of the LD phase between the 2 lines was assessed by calculating the LD correlation as a function of distance between SNP pairs (Figure 3). There appeared to be reasonable (correlation >0.80) LD phase persistence for SNP located within 200,000 bases. In gene detection exercises, this would probably be the distance providing sufficient persistence of the LD phase across the 2 lines. For genomic evaluation and selection purposes, results shown in Figure 3 imply that 200,000 to 300,000 SNP are needed to achieve persistent LD phase in the 2 lines. This is similar to the number proposed by de Roos et al. (2008) as the minimum required for achieving consistent LD and practicing genomic selection across divergent breeds.

In conclusion, slightly but statistically different LD values were observed in the 2 genetic lines. The group of cows sired by high-genetic-merit bulls was consistently in higher LD compared with control cows, suggestive of the potential role of selection in building LD. Analyzing each genetic line separately revealed stronger LD than looking at the entire data set. Given the limited experimental population size, independent confirmation of results would be desirable. Such results may be of value in gene detection and genome-wide association studies.

ACKNOWLEDGMENTS

The authors thank David Morrice (University of Edinburgh) for DNA extraction and dispatch, John Woolliams (University of Edinburgh) for haplotype confirmation software, Ross McGinn and Baggy Bagnall (Scottish Agricultural College) for collecting and storing blood, and the Scottish Government for funding.

REFERENCES

Clark, A. G. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. 7:111–122.

de Roos, A. P. W., B. J. Hayes, R. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein Friesian, Jersey and Angus cattle. Genetics 179:1503–1512.

Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10:381–391.

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome wide dense marker maps. Genetics 157:1819–1829.

Pritchard, J. K., and M. Przeworski. 2001. Linkage disequilibrium in humans: Models and data. Am. J. Hum. Genet. 69:1–14.

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer. 2008. Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Sci. 91:2106–2117.

Thévenon, S., G. K. Dayo, S. Sylla, L. Sidibe, D. Berthier, H. Legros, D. Boichard, A. Eggen, and M. Gautier. 2007. The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. Anim. Genet. 38:277–286.

Tier, B. 2006. Haplotyping for linkage disequilibrium mapping. Proc. 8th World Congr. Genet. Appl. Livest. Prod. CD-ROM Commun. 21–01.

Zhao, H., D. Nettleton, and J. C. M. Dekkers. 2007. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. Genet. Res. 89:1–6.