ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

doi:10.1111/j.1365-2052.2009.02011.x

# The pattern of linkage disequilibrium in German Holstein cattle

**S. Qanbari*, E. C. G. Pimentel*, J. Tetens†, G. Thaller†, P. Lichtner‡, A. R. Sharifi* and H. Simianer***

*Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, 37075 Göttingen, Germany. †Institute of Animal Breeding and Animal Husbandry, Christian-Albrechts-University, 24098 Kiel, Germany. ‡Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

## Summary

This study presents a second generation of linkage disequilibrium (LD) map statistics for the whole genome of the Holstein–Friesian population, which has a four times higher resolution compared with that of the maps available so far. We used DNA samples of 810 German Holstein–Friesian cattle genotyped by the Illumina Bovine SNP50K BeadChip to analyse LD structure. A panel of 40 854 (75.6%) markers was included in the final analysis. The pairwise $r^2$ statistic of SNPs up to 5 Mb apart across the genome was estimated. A mean value of $r^2 = 0.30 \pm 0.32$ was observed in pairwise distances of <25 kb and it dropped to $0.20 \pm 0.24$ at 50–75 kb, which is nearly the average inter-marker space in this study. The proportion of SNPs in useful LD ($r^2 \geq 0.25$) was 26% for the distance of 50 and 75 kb between SNPs. We found a lower level of LD for SNP pairs at the distance ≤100 kb than previously thought. Analysis revealed 712 haplo-blocks spanning 4.7% of the genome and containing 8.0% of all SNPs. Mean and median block length were estimated as $164 \pm 117$ kb and 144 kb respectively. Allele frequencies of the SNPs have a considerable and systematic impact on the estimate of $r^2$. It is shown that minimizing the allele frequency difference between SNPs reduces the influence of frequency on $r^2$ estimates. Analysis of past effective population size based on the direct estimates of recombination rates from SNP data showed a decline in effective population size to $N_e = 103$ up to ~4 generations ago. Systematic effects of marker density and effective population size on observed LD and haplotype structure are discussed.

**Keywords** bovine genome, effective population size, haplo-block structure.

## Introduction

Linkage disequilibrium (LD) defined as the non-random relationship between loci has recently been in the focus of attention. LD is the structural basis of 'Genomic Selection' programmes (Meuwissen *et al.* 2001) and helps to determine the actual genes responsible for variation of economically important traits (Van Laere *et al.* 2003; Grisart *et al.* 2004) through association mapping. The feasibility and efficiency of these approaches depends strongly on the extent, distribution and structure of LD, which determine how many markers are required for a genome scan in the population under study (Khatkar *et al.* 2007). Moreover, for high-resolution association mapping, it is also necessary to identify block-like structures of haplotypes and a minimal set of polymorphisms (haplotype tagging SNPs; htSNPs) that capture the most common haplotypes of each block (Johnson *et al.* 2001; Dawson *et al.* 2002). As a result of the variation in local recombination rates, mutation rates and genetic hitchhiking, the breakdown of LD is often discontinuous, producing haplotypic tracts across the genome (Ardlie *et al.* 2002; The International HapMap Consortium 2005). Simianer *et al.* (1997) demonstrated that this variability is also prevalent in the bovine genome and recombination probabilities even differ between families. As a result, today's chromosomes comprise a mosaic of haplotype blocks derived from ancestral chromosome fragments (e.g. Khatkar *et al.* 2007), and shared discrete haplotype blocks and LD patterns can be observed even in apparently unrelated individuals and populations (Gautier *et al.* 2007; Marques *et al.* 2008). Identifying these continental tracts can provide haplotypes to be used as genetic markers and

Address for correspondence

S. Qanbari, Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany.
E-mail: sqanbar@gwdg.de

*Accepted for publication 8 November 2009*

delimit regions where htSNPs can reasonably be defined. They could also provide information on the spacing of SNPs in association studies, i.e. where SNPs should be considered and where not. By adjusting for the differences in recombination rates across the genome, haplotype blocks can also be used for identifying the signatures of recent positive selection (Sabeti *et al.* 2002).

An increasing number of studies have aimed at quantifying LD characteristics in domestic animals, especially in cattle. Most of these studies used a low marker density or were performed in limited regions of the studied genomes. Farnir *et al.* (2000) performed the first whole-genome LD study to characterize the extent and pattern of LD based on the information of 284 microsatellite markers in Dutch Holstein cattle. Several subsequent studies have confirmed extensive LD in cattle (Khatkar *et al.* 2006a; Odani *et al.* 2006; Mckay *et al.* 2007; Marques *et al.* 2008; Nilsen *et al.* 2008; Prasad *et al.* 2008). They described an extensive LD and revealed that different measures of LD such as $r^2$ and $D'$ yield different conclusions in terms of the extent of LD. Recently, Sargolzaei *et al.* (2008) and Kim & Kirkpatrick (2009) reported a genome-wide LD profile based on the Affymetrix 10K SNP array in Holstein population of North America. They generally found a lower level of LD for SNP pairs than previously reported. Khatkar *et al.* (2007) reported a comprehensive genome-wide profile of LD statistics and haploblock characteristics based on a panel of 15 036 single nucleotide polymorphisms (SNPs) in Australian Holstein–Friesian cattle. The final average inter-marker spacing in their study was 251.8 kb, which is by a factor of $5 \times 10^{-3}$ less dense than the panel currently being used in LD analysis of human genome. However, it is now known that the BTAu_3.1 build used to physically locate SNPs in their study has inconsistencies with other independently built cattle maps (Marques *et al.* 2007; Snelling *et al.* 2007). More recently, Villa-Angulo *et al.* (2009) used a panel of 31 857 SNPs generated by the Bovine HapMap Consortium to characterize a high-resolution haplotype block structure of 19 breeds of different geographic origin. They focused mainly on 101 high density regions spanning up to 7.6 Mb on three chromosomes 6, 14 and 25 with an average density of approximately one SNP per 4 kb.

With the availability of larger-scale SNP data sets, it has become possible to construct LD maps with higher resolution. In this study, we use SNP data generated with the Illumina Bovine SNP50K BeadChip to create a second generation LD map of Holstein–Friesian cattle. We also explore some properties of $r^2$ as the most common measure of LD in this study.

## Materials and methods

### Data preparation and haplotype reconstruction

The subset of animals used in this study is part of the total population of Holstein cattle genotyped for the genomic selection programme in Germany. Semen or blood samples from 810 German Holstein–Friesian cattle including 469 bulls and 341 bull dams were used as the source of genomic DNA. DNA was purified applying a modified protocol according to Miller *et al.* (1988) including an additional dithiothreitol-treatment for the semen samples and genotyped using the Illumina Bovine SNP50K BeadChip (Matukumalli *et al.* 2009). This chip contains a total of 54 001 SNPs with a mean neighbour marker distance of 48.75 kb. 1728 SNP loci were excluded because of unknown genomic position and 11 markers were monomorphic. For the purposes of this study, only autosomal SNPs with minor allelic frequencies (MAF) ≥0.05 were included in the LD analysis. The number of heterozygous loci was determined and used to estimate the average heterozygosity for all individuals. The allele frequencies, observed heterozygosity and expected heterozygosity for each SNP were determined.

For this analysis, fully phased haplotype data were required. As a result of the considerable number of animals analysed, we disregarded the probable effect of founders and assumed animals to reflect a representative sample of the currently relevant breeding population. Both paternal and maternal haplotypes were utilized for the estimation of LD. After the aforementioned filtering process, we reconstructed haplotypes for each chromosome using default options in fastPHASE (Scheet & Stephens 2006).

### Measure of LD

Several statistics have been used to measure the LD between a pair of loci. The two most common measures are the absolute value of $D'$, and $r^2$, both derived from Lewontin's $D$ (Lewontin 1964). We used $r^2$, which is generally accepted as the more robust and better interpretable LD parameter (Kruglyak 1999; Ardlie *et al.* 2002; Terwilliger *et al.* 2002).

Consider two loci, A and B, each locus having two alleles (denoted $A_1$, $A_2$; $B_1$, $B_2$ respectively). We denote $f_{11}$, $f_{12}$, $f_{21}$ and $f_{22}$ as the frequencies of the haplotypes $A_1B_1$, $A_1B_2$, $A_2B_1$ and $A_2B_2$ respectively; $f_{A_1}$, $f_{A_2}$, $f_{B_1}$ and $f_{B_2}$ are the frequencies of $A_1$, $A_2$, $B_1$ and $B_2$ respectively; following Hill and Weir (1994):

$$r^2 = \frac{(f_{11}f_{22} - f_{12}f_{21})^2}{f_{A_1}f_{A_2}f_{B_1}f_{B_2}}.$$

### LD haplo-block partitioning

Existing block definition algorithms are based on two alternative methods: Either pairwise $D'$ values above a lower limit are used to detect regions of little or no recombination (Daly *et al.* 2001; Gabriel *et al.* 2002; Wang *et al.* 2002), or blocks are defined by employing some haplotypic diversity criterion, where a small number of common haplotypes provide high chromosomal frequency coverage (Patil *et al.* 2001; Zhang *et al.* 2002, 2003; Anderson & Novembre


doi:10.1111/j.1365-2052.2009.02011.x

2003). For the purpose of this study, we used the algorithm suggested by Gabriel *et al.* (2002) defining a pair of SNPs to be in 'strong LD' if the upper 95% confidence bound of D′ is between 0.7 and 0.98. Reconstructed haplotypes were inserted into HAPLOVIEW v4.1 (Barrett *et al.* 2005) to estimate LD statistics and construct the blocking pattern as well as identify haplotype-tagging SNPs for all 29 autosomes.

### Estimating effective population size using LD

According to Wright (1938), effective population size ($N_e$) is defined as 'the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration'. $N_e$ provides useful information about the population evolution and improves the understanding and modelling of the genetic architecture underlying complex traits (Reich & Lander 2001). $N_e$ can be estimated from LD data and the availability of dense markers has made this option feasible. Sved (1971) has formulated the relationship of LD and $N_e$ in the absence of mutation as $r^2 = 1/(4N_ec + 1)$, where $c$ represents the linkage map distance in Morgans. If mutation is accounted for in the model, the expectation of $r^2$ is $1/(4N_tc + 2)$, where $N_t$ is the population size $1/(2c)$ generations ago. For more information, we refer to Tenesa *et al.* (2007). In this study, we assessed genetic distance $c$ directly by estimating the recombination rates across the genome using PHASE V.2.1 (Li and Stephens 2003). For this purpose, random segments of 15 Mb were selected on each autosome. The recombination model was applied based on 100 individuals and increasing the number of iterations of the final run 10 times to obtain better estimates of uncertainty. The prior value for effective population size was set to 100. To save computing time, we used known haplotypes with fragment sizes of 12 bp. An average of $N_e$ over chromosomes was then calculated corresponding to the various times in the past. We inferred $N_e$ for each autosomal chromosome at distance bins of <0.025, 0.025–0.05, 0.05–0.1, 0.1–0.5, 0.5–1, 1–5 and 5–15 cM. This range of linkage map distance infers the past effective population size up to 2000 generations ago.

## Results

### Marker statistics and genetic diversity

A total of 40 854 (75.65%) markers passed the above filtering criteria and were included into the final analysis. This subset of markers covers 2544.1 Mb of the genome with 62.27 ± 58.3 kb average adjacent marker spacing. The largest gap between SNPs (2081.5 kb) was located on chromosome 10. For the SNPs analysed in this study, the average observed heterozygosity and mean MAF were estimated as 0.37 ± 0.12 and 0.28 ± 0.15 respectively. The SNPs genotyped showed an almost uniform distribution across frequency classes. This is probably caused by the construction of the SNP array, which was optimized with respect to a uniform SNP spacing and MAF distribution. The observed heterozygosity in the studied Holstein population averaged 0.23.

### Pattern of haplotype blocks

The identification of haplotype blocks and the minimal set of htSNPs required to capture haplotype variation in a population sufficiently is critical for association studies, and will reduce cost and effort. Table 1 presents a descriptive summary of genome-wide marker and haplo-block distribution in the data set analysed. A total of 712 haplo-blocks spanning 118 859 kb (4.67%) of the genome were detected. Mean and median block length were estimated as 164 ± 117 and 144 kb respectively with a maximum of 1261 kb. The distribution of haplotype block size is depicted in Fig. 1. Chromosome 1, having 57 blocks spanning 9159 kb and Chromosome 27 with 3 blocks covering 408 kb, showed the longest and the shortest haplotypic structures in the genome. In total, 3258 SNPs (7.97% of all SNPs used) formed blocks with a range of 2–11 SNPs per tract. Using the tagger option incorporated in HAPLOVIEW, 36 301 SNPs (89% of all used SNPs) were tagged in the data set analysed. These SNPs can tag either neighbouring markers or a set of common haplotypes within an LD block. Figure 2 displays the distribution of htSNPs across the genome of the studied population. The ratio htSNP/nSNP shows a negative association with the number of SNPs reflecting chromosome length.

### Extent of LD across the genome

All possible SNP pairs with distance ≤5 Mb on the same chromosome produced 3 216 038 pairwise LD values on the 29 bovine autosomes. To visualize the decay of LD and the proportion of pair markers in useful LD, we stacked $r^2$ values and plotted them as a function of inter-marker distance categories (<0.025, 0.025–0.05, 0.05–0.075, 0.075–0.12, 0.12–0.2, 0.2–0.5, 0.5–1.5, 1.5–3 and 3–5 (Mb) (Fig. 3). This genome-wide bar plot illustrates the rate at which LD decays with physical distance and forms the basis for comparison between studies. We observed an inverse relationship between LD and marker distance, confirming recent studies on $r^2$ measures in cattle. Overall, six cases of complete LD ($r^2 = 1.0$) were observed for the entire genome.

The mean $r^2$ values and the proportion of SNP pairs that show statistically significant LD for SNP pairs up to 5 Mb apart are presented in Table 2. A mean value of $r^2 = 0.30 \pm 0.32$ was observed in pairwise distances of <25 kb and it dropped to 0.20 ± 0.24 at 50–75 kb, the interval which includes the average inter-marker space in


doi:10.1111/j.1365-2052.2009.02011.x

**Table 1** Genome-wide summary of marker and haplotype blocks in the Holstein cattle.

| Chr | Initial (*n*) | Final (*n*) | Chr-length (Mb) | Linkage map (cM) | Block (*n*) | Block-length (kb) | Mean-BL ± SD (kb) | BSNPs[1] (*n*) | htSNPs (*n*) | Max gap (kb) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3343 | 2641 | 161.1 | 154 | 57 | 9159 | 160.7 ± 112 | 267 | 2263 | 683.9 |
| 2 | 2764 | 2149 | 140.6 | 126 | 35 | 6181 | 176.6 ± 146 | 167 | 1876 | 651.9 |
| 3 | 2566 | 2037 | 127.9 | 128 | 52 | 7428 | 142.8 ± 102 | 216 | 1790 | 813.7 |
| 4 | 2541 | 1999 | 124.1 | 119 | 41 | 7173 | 175.0 ± 118 | 197 | 1759 | 889.7 |
| 5 | 2181 | 1718 | 125.8 | 135 | 30 | 6333 | 211.1 ± 147 | 149 | 1521 | 1050.5 |
| 6 | 2535 | 2044 | 122.5 | 134 | 46 | 7918 | 172.1 ± 119 | 225 | 1778 | 826.2 |
| 7 | 2294 | 1767 | 112.1 | 135 | 34 | 6919 | 203.5 ± 127 | 177 | 1519 | 657.0 |
| 8 | 2362 | 1849 | 116.9 | 128 | 42 | 7586 | 180.6 ± 108 | 196 | 1588 | 738.3 |
| 9 | 2036 | 1623 | 108.1 | 116 | 20 | 3879 | 194.0 ± 136 | 89 | 1469 | 760.8 |
| 10 | 2179 | 1713 | 106.2 | 118 | 40 | 4787 | 119.7 ± 93 | 166 | 1519 | 2081.5 |
| 11 | 2267 | 1813 | 110.2 | 130 | 27 | 4658 | 172.5 ± 104 | 126 | 1624 | 989.5 |
| 12 | 1683 | 1320 | 85.3 | 109 | 20 | 3102 | 155.1 ± 154 | 85 | 1190 | 788.7 |
| 13 | 1802 | 1396 | 84.3 | 105 | 32 | 4793 | 149.8 ± 93 | 136 | 1227 | 608.9 |
| 14 | 1722 | 1356 | 81.3 | 103 | 28 | 5402 | 192.9 ± 96 | 141 | 1166 | 576.0 |
| 15 | 1688 | 1365 | 84.6 | 109 | 19 | 3157 | 166.2 ± 116 | 81 | 1245 | 660.2 |
| 16 | 1606 | 1251 | 77.8 | 94 | 31 | 6443 | 207.8 ± 274 | 151 | 1087 | 1015.4 |
| 17 | 1585 | 1284 | 76.5 | 95 | 14 | 1971 | 140.8 ± 68 | 64 | 1170 | 840.4 |
| 18 | 1351 | 1100 | 66.1 | 84 | 12 | 1635 | 136.3 ± 54 | 53 | 1012 | 896.4 |
| 19 | 1378 | 1108 | 65.2 | 109 | 15 | 2876 | 191.7 ± 96 | 77 | 1006 | 553.1 |
| 20 | 1564 | 1252 | 75.7 | 82 | 23 | 3713 | 161.4 ± 89 | 102 | 1099 | 837.1 |
| 21 | 1419 | 1093 | 69.2 | 83 | 16 | 2279 | 142.4 ± 102 | 65 | 985 | 849.4 |
| 22 | 1299 | 1009 | 61.8 | 88 | 15 | 1723 | 114.9 ± 85 | 58 | 903 | 601.3 |
| 23 | 1083 | 871 | 53.3 | 80 | 7 | 1500 | 214.3 ± 150 | 35 | 805 | 476.3 |
| 24 | 1294 | 1013 | 64.9 | 78 | 13 | 1944 | 149.5 ± 113 | 56 | 916 | 527.3 |
| 25 | 987 | 810 | 44.0 | 68 | 15 | 1834 | 122.3 ± 94 | 68 | 752 | 589.9 |
| 26 | 1086 | 849 | 51.7 | 79 | 12 | 2136 | 178.0 ± 180 | 48 | 763 | 682.6 |
| 27 | 977 | 798 | 48.7 | 67 | 3 | 408 | 136.0 ± 46 | 13 | 748 | 1776.8 |
| 28 | 942 | 779 | 46.0 | 61 | 3 | 461 | 153.7 ± 156 | 12 | 740 | 470.6 |
| 29 | 1048 | 847 | 52.0 | 69 | 10 | 1461 | 146.1 ± 104 | 38 | 781 | 1505.8 |
| Total | 51′582 | 40′854 | 2544.1 | 2986 | 712 | 118′859 | 164.4 ± 117 | 3258 | 36′301 | 2081.5 |

[1]Number of SNPs forming haplo-blocks.

this study. In contrast, an overall mean value of $r^2 = 0.21 \pm 0.26$ was observed for SNPs less than 100 kb apart from each other compared with $r^2 = 0.59$ presented by Sargolzaei *et al.* (2008) for the north American Holstein cattle. A similar study by Kim & Kirkpatrick (2009) revealed strong LD ($r^2 > 0.8$) in genomic regions of approximately 50 kb or less, which is much larger than the observation of this study ($r^2 = 0.29$).

The threshold for useful LD that was chosen in this study is 0.25. With this threshold, and considering that on average 1 cM is equivalent to 1 Mb, useful LD extended over 0.5–1.5 cM so that the proportion of SNP pairs in useful LD is above 5%. The proportion of SNPs in useful LD was 39% for the distance of 25 kb or less between SNPs. This proportion dropped to 0.26% for SNPs between 50 and 75 kb apart from each other. Overall, for SNPs less than 100 kb apart from each other the proportion of SNPs in useful LD was 29%. This proportion was reported as 68.34% by Sargolzaei *et al.* (2008) even with a higher threshold (0.3). However, the substantial LD estimated for SNP pairs more than 100 kb apart ($r^2 = 0.14$) is similar.

It is known that LD between SNPs with a low minor allele frequency is biased upwards, and thus high-frequency polymorphisms are preferable for accurate estimation of LD (Reich *et al.* 2001). In part, this can be explained by statistical properties of the LD statistics (Dunning *et al.* 2000), but it may also have an evolutionary interpretation because low frequency SNPs have a higher probability of having arisen recently (Nordborg & Tavaré 2002). Taking this into account, we evaluated the decay of LD for the SNPs with MAF greater than 15% to elucidate its utility in terms of having SNP pairs in useful LD for genomic association analysis. We observed an increase of about 10% in frequency of SNP pairs representing useful LD for almost all physical distance bins up to 5 Mb (Fig. 4).

## LD properties

The decay of LD measures with increasing physical distance is well documented. LD is expected to be a function of linkage distance in animal populations, at least for tightly linked loci. It is also reported that SNPs of divergent MAFs


doi:10.1111/j.1365-2052.2009.02011.x

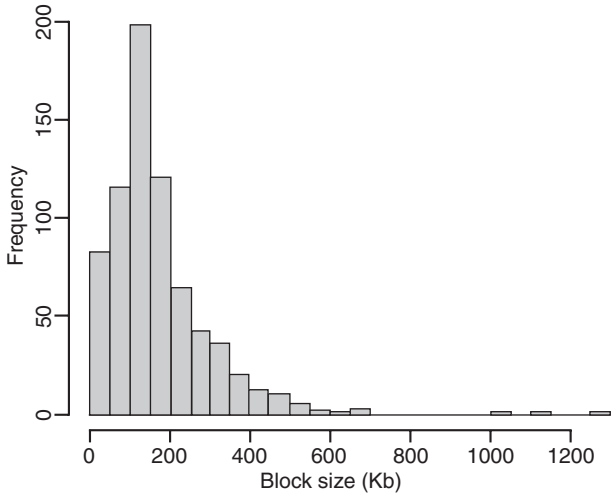


**Figure 1** Distribution of haplo-block size in the Holstein cattle genome.

on average have different LD properties (Pritchard & Przeworski 2001). Figure 5 displays the decay of LD as a function of physical distance and absolute MAF difference (ΔMAF) between SNP pairs. It can be seen that pairwise $r^2$ decreases with increasing distance and increasing ΔMAF. It is evident that the dependence of $r^2$ on distance is stronger than its dependence on ΔMAF. It is also shown that SNP pairs in short physical distance are more affected by ΔMAF. The magnitude of this dependency in the case of SNP pairs far from each other is negligible.

To explore the dependence of LD on allele frequency, we calculated the average $r^2$ statistic within nine bins of physical distance between frequency-matched pairs of SNPs with ΔMAF ≤10% and compared results with the average $r^2$ between all SNP pairs (Table 2). Mean $r^2$ values were higher between matched SNP pairs than between non-matched ones for all distance bins, with a difference of around 50% in the shortest distances. For the markers within a distance range of 50–75 kb, the proportion of SNP pairs in useful LD increased from 26% to 39%. We observed a higher extent of LD for frequency-matched vs. non-matched pairs of SNPs. As such, with frequency-matched pairs of SNPs, LD significantly extended up to the range of 1.5–3 Mb.

In a further step, we plotted the $r^2$ vs. minor allele frequencies of both loci (Fig. 6). SNP pairs with the highest MAF interval represent the lowest $r^2$ and *vice versa.* Frequency-matched SNPs with moderate or low MAF values both result in the highest $r^2$ regions. However, there is a trend demonstrating a slight raise of LD for matched SNPs with moderate MAF compared with the matched SNPs with lower MAF. Therefore, it can be concluded that the frequency matched SNP pairs are less influenced when calculating pairwise $r^2$ values, substantiating lower decay of LD for these loci.

## Past effective population size

In most studies so far, genetic distance $c$ was approximated by using physical distance directly (1 Mb–1 cM) for the estimation of $N_e$ (Gautier *et al.* 2007; Hayes *et al.* 2008; Kim & Kirkpatrick 2009). In this study, we estimated the recombination rates directly from dense SNP data. Figure 7 displays the decay of LD as a function of recombination rate between pairs of SNPs. Recombination rates are not constant within chromosomes and vary among regions. Overall, a correlation of −0.22 was observed between $r^2$ and recombination rate over all adjacent marker intervals analysed. LD values averaged in bins of estimated linkage distance were used to study the changes in effective population size of the population from 2000 generations ago up to the present. We compared the results with the estimates of *Ne* based on available cattle linkage map information (http://www.marc.usda.gov/genome/cattle/cattle.html). Given the known linkage and physical lengths of chromosomes (Table 1), we transformed the physical position

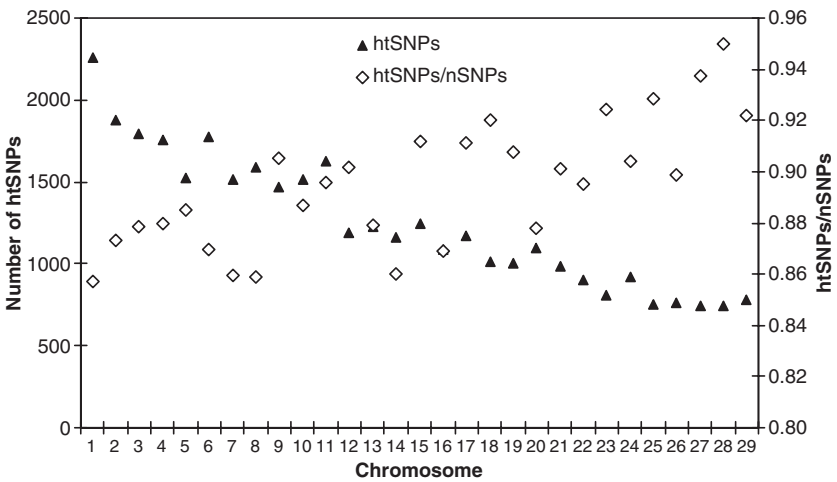


**Figure 2** Distribution of htSNPs across the genome of Holstein population studied. Triangles display the number of htSNPs for each chromosome and diamonds represent the ratio of htSNPs vs. SNPs analysed for each chromosome.


doi:10.1111/j.1365-2052.2009.02011.x

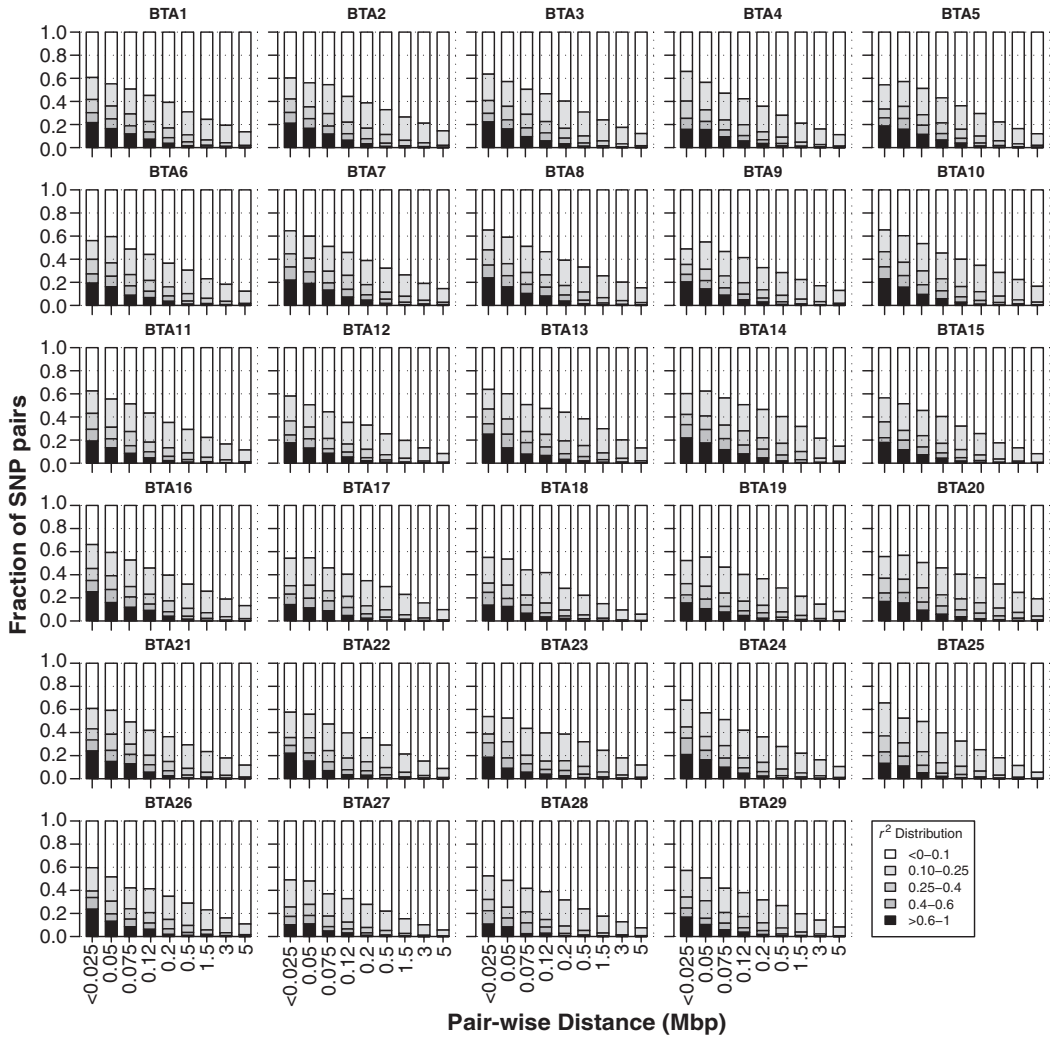


**Figure 3** Level of linkage disequilibrium decay as a function of distance between pairs of SNPs up to 5 Mb for the entire genome.

**Table 2** Frequency and mean $r^2$ estimated for SNP pairs in different distances compared with the frequency matched SNP pairs.

| Distance (Mb) | SNP pairs (*n*) | | Median $r^2$ | | Mean $r^2$ ± SD | | Frequency $r^2 \geq 0.25$ (%) | |
|---|---|---|---|---|---|---|---|---|
| | All pairs | ΔMAF ≤ 0.1 | All pairs | ΔMAF ≤ 0.1 | All pairs | ΔMAF ≤ 0.1 | All pairs | ΔMAF ≤ 0.1 |
| <0.025 | 6002 | 4617 | 0.16 | 0.39 | 0.30 ± 0.32 | 0.45 ± 0.38 | 39 | 56 |
| 0.025–0.05 | 20 108 | 12 735 | 0.13 | 0.25 | 0.25 ± 0.28 | 0.38 ± 0.35 | 34 | 50 |
| 0.05–0.075 | 17 938 | 8340 | 0.09 | 0.14 | 0.20 ± 0.24 | 0.29 ± 0.31 | 26 | 39 |
| 0.075–0.12 | 31 833 | 10 725 | 0.07 | 0.09 | 0.16 ± 0.20 | 0.22 ± 0.27 | 20 | 30 |
| 0.12–0.2 | 55 778 | 12 906 | 0.06 | 0.06 | 0.12 ± 0.16 | 0.16 ± 0.22 | 15 | 22 |
| 0.2–0.5 | 204 584 | 28 572 | 0.04 | 0.04 | 0.09 ± 0.12 | 0.11 ± 0.16 | 10 | 15 |
| 0.5–1.5 | 664 447 | 52 743 | 0.03 | 0.03 | 0.07 ± 0.09 | 0.08 ± 0.12 | 6 | 9 |
| 1.5–3 | 965 989 | 35 720 | 0.02 | 0.02 | 0.05 ± 0.07 | 0.06 ± 0.09 | 3 | 5 |
| 3–5 | 1 249 359 | 17 384 | 0.02 | 0.02 | 0.04 ± 0.06 | 0.05 ± 0.07 | 1 | 3 |

to the approximate linkage distance between pairs of SNPs and averaged the estimates over chromosomes. While $N_e$ was inferred as 1113 for 500 generations ago, estimates based on recombination rates show a decline in effective population size to 103 up to ~4 generations ago (Fig. 8a). This is close to the estimate ($N_e \leq 100$) in the North American Holstein population based on the analysis of both LD (Kim & Kirkpatrick 2009) and


doi:10.1111/j.1365-2052.2009.02011.x

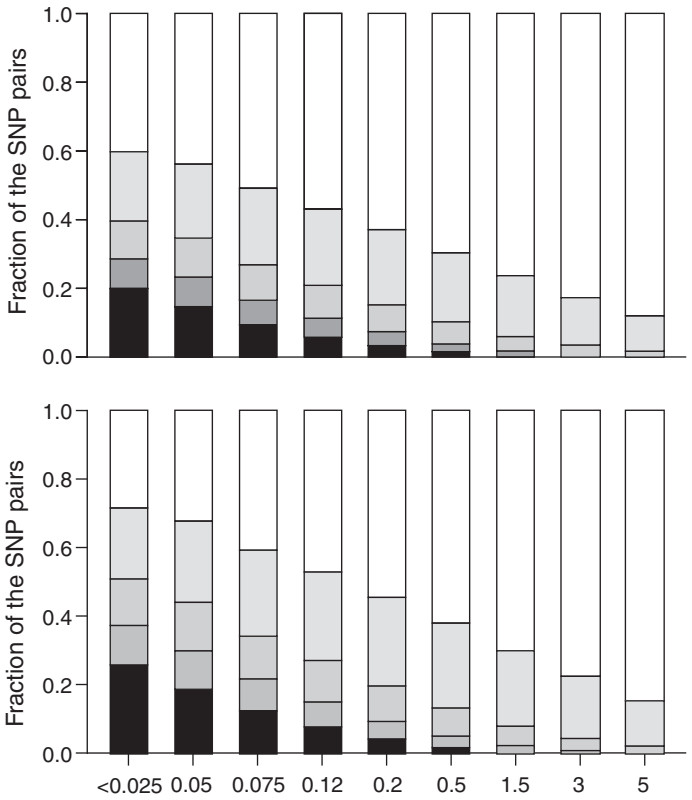


**Figure 4** Comparison of fraction of marker pairs with different $r^2$ levels (<0.1, ≥0.25, ≥0.4, ≥0.6 and >0.6, depicted by different colours) for marker pairs in different distance bins maximum 5 Mb. (a) SNP pairs of all 40 854 SNPs with minor allelic frequencies (MAF) ≥5%; (b) considering only SNP pairs with MAF ≥0.15.

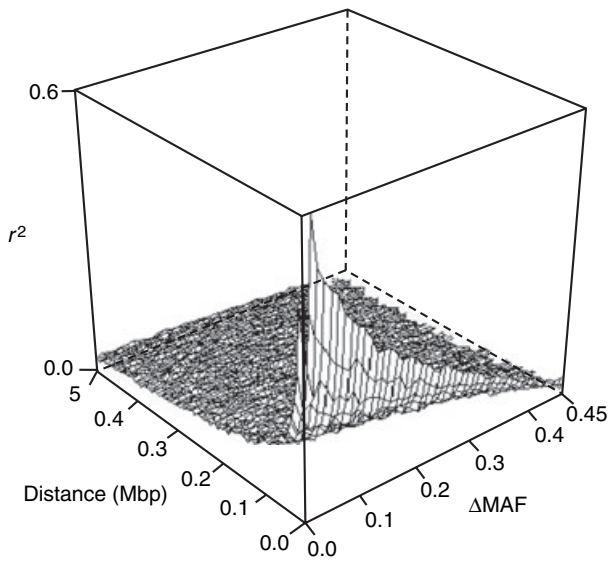


**Figure 5** Three-dimensional surface plot depicting the decay of linkage disequilibrium vs. inter-marker distance and minor allelic frequency interval.

inbreeding rate (Young & Seykora 1996). With the mutation included model, it drops to 56, which is close to the inbreeding-based estimates of $N_e$ < 50 and $N_e$ = 52 in the Danish (Sorensen *et al.* 2005) and German (Koenig & Simianer 2006) Holstein populations respectively.

## Discussion

Linkage disequilibrium maps increase power and precision in association mapping, define optimal marker spacing, and identify recombination hot-spots and regions influenced by natural selection. In this report, we present an analysis of LD of 40 854 markers densely distributed across the entire bovine genome in a sample of German Holstein cattle. Although the principle of LD is fairly simple (i.e. the non-random segregation of markers in close proximity), the complex interplay between all confounding factors complicates the interpretation of LD results. As LD depends on the age of the SNP-creating mutations, the demographic population history, genetic drift, the recombination fraction, directional selection, population stratification and other factors, it is highly variable even between close loci (Kruglyak 1999; Ardlie *et al.* 2002; Pritchard & Przeworski 2001). As a result, two markers that are very close together can exhibit a low level of LD, while markers that are more distant can show a higher than expected level of LD.

In this study, we used pairwise $r^2$ statistics up to 5 Mb across the bovine genome to estimate the extent of LD. The first reports on the extent of LD in cattle genome described a long range of LD (e.g. up to 20 cM) (Farnir *et al.* 2000; Tenesa *et al.* 2003). Further analyses with denser markers confirmed extensive LD, but in general found lower levels (Spelman & Coppieters 2006, Khatkar *et al.* 2006a). Recently, two genome-wide studies based on 10 K SNP data


doi:10.1111/j.1365-2052.2009.02011.x

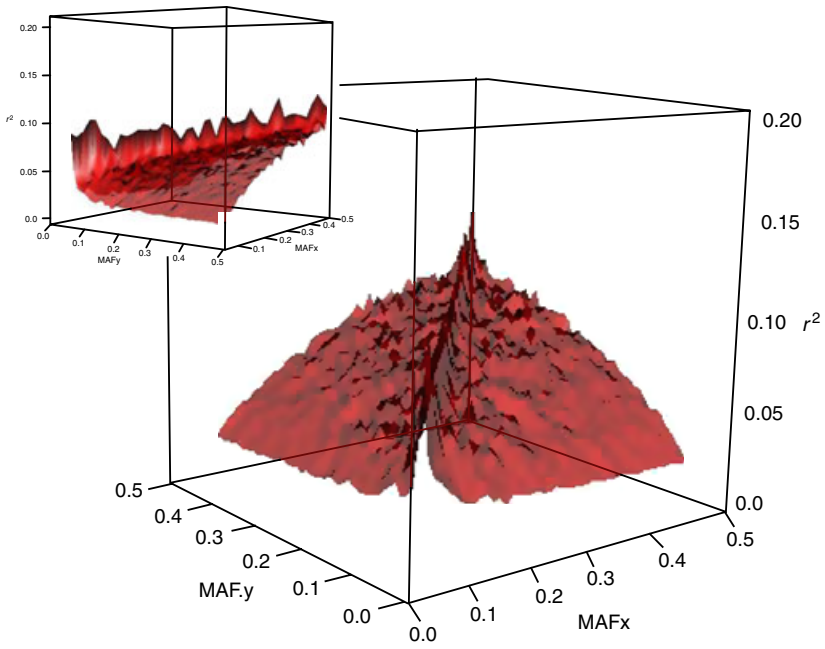


**Figure 6** Prospective plot depicting the decay of linkage disequilibrium with allele frequencies of SNP pairs. $r^2$ means were calculated for 45 bins of 0.01 allele frequency each.

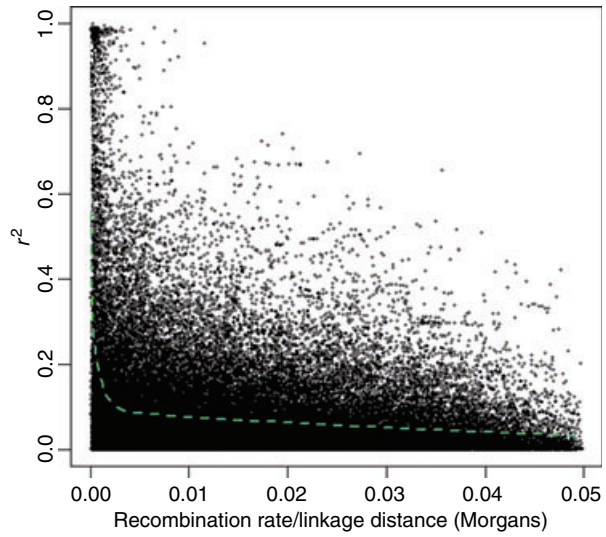


**Figure 7** Linkage disequilibrium between SNP pairs was plotted on the estimates of recombination rate as a measure of linkage distance (M).

have revealed that the level of LD is less than previously thought (Sargolzaei *et al.* 2008; Kim & Kirkpatrick 2009). The results of this study demonstrate even less LD for SNP pairs at distances ≤100 kb.

It was suggested that LD within genes is higher than LD in inter-genic regions, at least for tightly linked markers (Kim & Kirkpatrick 2009), hence the discrepancy observed may be attributed to a systematic difference of the selected set of SNPs. For the Illumina Bovine SNP50k BeadChip, SNPs were mainly selected to evenly cover the entire genome, while in other studies the SNPs were targeted to certain candidate regions. The average LD over the entire genome is the quantity of interest, especially for use in genomic selection and whole-genome association mapping without prior positional information, which was evaluated in our study. In general, it is difficult to compare the level of LD obtained in different studies because of different sample sizes, LD measures, marker types, marker densities and recent and historical population demographics (Pritchard & Przeworski 2001).

The decay of LD in a genome determines the power of QTL detection in association mapping studies and indicates the required marker density. It was shown that in indirect association studies, the sample size must be increased by roughly $1/r^2$ when compared with the sample size for detecting the causal mutation directly (Kruglyak 1999; Pritchard & Przeworski 2001). Meuwissen *et al.* (2001) simulated the required level of LD ($r^2$) for genomic selection to achieve an accuracy of 0.85 for genomic breeding values to be 0.2. Ardlie *et al.* (2002) defined high values of LD as $r^2 > 1/3$. In this study, we assumed the threshold of useful LD to be 0.25. To achieve this level, our results indicate that the SNP spacing should be ∼35 kb in future population-wide studies with a whole-genome approach. This implies the use of more than 75 000 SNPs per individual, assuming that all SNPs are informative (with a MAF ≥ 0.05). According to the results of this study, the same power can be achieved by implementing a panel of 50 000 SNPs with moderate frequencies (e.g. MAF ≥ 0.15), which simultaneously improves the accuracy and magnitude of estimated LD between pairs of SNPs.

In this study, we examined the decay of LD as a function of physical distance. Despite the LD map showing a distinct decrease of LD values over increasing physical distance, the LD also showed extensive variability between genomic regions and chromosomes. This variation was probably


doi:10.1111/j.1365-2052.2009.02011.x

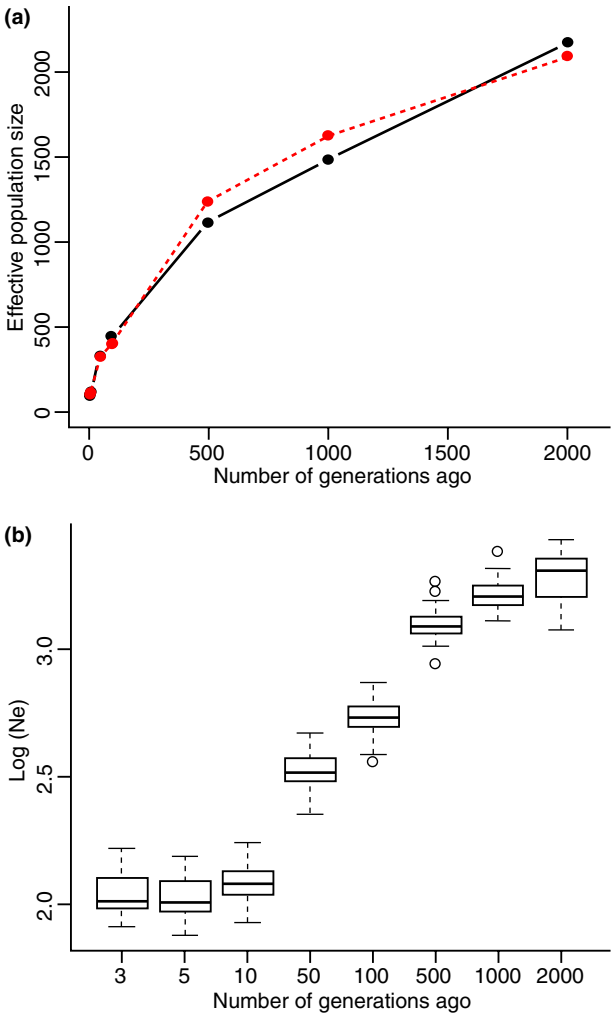


**Figure 8** Estimated effective population size over the past generations from linkage disequilibrium data. (a) Dashed and solid lines represent $N_e$ based on estimates of recombination rates and approximate linkage distances respectively. (b) Boxplot representing the trend of $\log_{10}(N_e)$ over time. The variability at each point of time reflects the variation of estimates between the 29 autosomes.

attributed to recombination rates varying between and within chromosomes, heterozygosity, genetic drift and effects of selection.

The impact of allele frequency in analysing genome-wide LD was also explored in this study. Our results demonstrate that the dependence of LD on the MAF interval of SNP pairs is stronger for SNPs separated by short distances. These results also reveal that minimizing the allele frequency difference between SNPs provides a more sensitive and useful metric for analysing LD across the bovine genome. Although an entirely frequency-independent measure of LD is not possible (Lewontin 1988), frequency matching between SNP pairs removes one major source of statistical noise when assessing the LD structure.

There are several published studies reporting LD properties based on dense SNP markers in cattle. Khatkar *et al.* (2006a) pioneered exploiting dense SNPs in developing bovine LD maps by characterizing the LD profile for chromosome 6 in the Australian Holstein population. They used 220 SNPs and confirmed an extensive level of LD in Holstein cattle. Gautier *et al.* (2007) studied LD properties in cattle breeds of different origin and observed that the haplotype blocks extended up to 700 kb in some cattle breeds. Khatkar *et al.* (2007) developed a primary genome-wide LD map based on a panel of 9195 informative SNPs, reporting 727 blocks with three or more SNPs (mean length = 69.7 kb) covering 2.18% of the genome. In a similar study, Marques *et al.* (2008) compared LD properties of chromosome 14 in Holstein and Angus cattle and reported 64 blocks (33 bp–1126 kb). Recently, Kim & Kirkpatrick (2009) reported 119 haplo-blocks (with more than 4 SNPs) with a mean length of 26.2 kb in a whole-genome scan of Holstein cattle. It was suggested that as the number of markers increases more haplotype blocks will be identified. This was confirmed by Villa-Angulo *et al.* (2009), who reported blocks of smaller size with an overall mean of 10.3 kb across 19 breeds. However, compared with the marker density used in the previous studies, this study, which used more SNPs and reported 712 blocks does not follow this expectation. Similar to the LD differences observed, this could also be caused by the use of a different set of markers, which are evenly distributed across the genome covering both genic and inter-genic regions. Although the number of blocks is not different, we observed a higher block coverage percentage compared with the Australian population.

The average extent of LD in the human genome has been extensively studied: it extends a few kb up to 50 kb but is highly variable, depending on the population and threshold used to measure LD. For the bovine genome, Villa-Angulo *et al.* (2009) reported a similar block size to that observed in the human genome. However, this was limited to targeted regions of the genome on three chromosomes, characterized by a high SNP density. Compared with the results of human studies, average block sizes observed in this study on the bovine genome are 20–30 times larger than haplotype blocks found in the human genome (Hinds *et al.* 2005). It must be noted that the marker density used in this study is about 100 times sparser than the one currently being used for the human genome. Hence, some of the long blocks observed in this study may break down to smaller tracts if the SNP density was increased. However, as a result of the smaller effective population size of cattle compared with the human population (Hayes *et al.* 2003) and the relatively high inbreeding frequency, a greater level of LD and larger haplotype blocks are expected.

## Conclusions

We present a second generation of LD map statistics for the Holstein genome, which has four times higher resolution compared with that of the maps available so far. We found a


doi:10.1111/j.1365-2052.2009.02011.x

lower level of LD for SNP pairs at distances ≤100 kb than previously reported. Assuming that $r^2 > 0.25$ is useful for association studies, the level of LD obtained in this study indicates that a denser SNP map would be beneficial to capture the LD information required for whole-genome fine-mapping and genomic selection and to completely assess the pattern of LD across the genome. The results show that frequency matched SNP pairs reduce the dependence of $r^2$ on allele frequency and provide a useful metric for analysing LD. The larger block size in Holstein cattle observed in this study indicates substantially greater LD in cattle than in human populations.

## Acknowledgements

This study is part of the project FUGATO-plus GenoTrack and was financially supported by the German Ministry of Education and Research, BMBF, the Förderverein Biotechnologieforschung e.V. (FBF), Bonn, and Lohmann Tierzucht GmbH, Cuxhaven. SQ thanks the H. Wilhelm Schaumann Stiftung Hamburg for financial support.

## References

Anderson E. & Novembre J. (2003) Finding haplotype block boundaries by using the minimum-description-length principle. *American Journal of Human Genetics* **73,** 336–54.

Ardlie K.G., Kruglyak L. & Seielstad M. (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* **3,** 299–309.

Barrett J.C., Fry B.J., Maller J. & Daly M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–5.

Daly M.J., Rioux J.D., Schaffner S.F., Hudson T.J. & Lander E.S. (2001) High-resolution haplotype structure in the human genome. *Nature Genetics* **29,** 229–32.

Dawson E., Abecasis G.R., Bumpstead S. *et al.* (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418,** 544–8.

Dunning A.M., Durocher F., Healey C.S. *et al.* (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics* **67,** 1544–54.

Farnir F., Coppiettiers W., Arranz J.J. *et al.* (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10,** 220–7.

Gabriel S.B., Schaffner S.F., Nguyen H. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science* **296,** 2225–9.

Gautier M., Faraut T., Moazami-Goudarzi K. *et al.* (2007) Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics* **177,** 1059–70.

Grisart B., Farnir F., Karim L. *et al.* (2004) Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 2398–403.

Hayes B.J., Visscher P.M., McPartlan H.C. & Goddard M.E. (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research* **13,** 635–43.

Hayes B.J., Lien S., Nilsen H. *et al.* (2008) The origin of selection signatures on bovine chromosome 6. *Animal Genetics* **39,** 105–11.

Hill W.G. & Weir B.S. (1994) Maximum likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics* **54,** 705–14.

Hinds D.A., Stuve L.L., Nilsen G.B. *et al.* (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* **307,** 1072–9.

Johnson G.C., Esposito L., Barratt B.J. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29,** 233–7.

Khatkar M.S., Collins A., Cavanagh J.A.L. *et al.* (2006a) A first generation metric linkage disequilibrium map of bovine chromosome 6. *Genetics* **174,** 79–85.

Khatkar M.S., Zenger K.R., Hobbs M. *et al.* (2007) A primary assembly of a bovine haplotype block map based on a 15,036 single nucleotide polymorphism panel genotyped in Holstein Friesian cattle. *Genetics* **176,** 763–72.

Kim E.S. & Kirkpatrick B.W. (2009) Linkage disequilibrium in the North American Holstein population. *Animal Genetics* **40,** 279–88.

Koenig S. & Simianer H. (2006) Approaches to the management of inbreeding and relationship in the German Holstein dairy cattle population. *Livestock Science* **103,** 40–53.

Kruglyak L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22,** 139–44.

Lewontin R.C. (1964) The interaction of selection and linkage I. General considerations; heterotic models. *Genetics* **49,** 49–67.

Lewontin R.C. (1988) On measures of gametic disequilibrium. *Genetics* **120,** 849–52.

Li N. & Stephens M. (2003) Modelling Linkage Disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* **165,** 2213–33.

Marques E., De Givry S., Stothard P. *et al.* (2007) A high resolution radiation hybrid map of bovine chromosome 14 identifies scaffold rearrangement in the latest bovine assembly. *BMC Genomics* **8,** 254.

Marques E., Schnabel R., Stothard P. *et al.* (2008) High density linkage disequilibrium maps of chromosome 14 in Holstein and Angus cattle. *BMC Genetics* **9,** 45.

Matukumalli L.K., Lawley C.T., Schnabel R.D. (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* **4,** e5350.

Mckay S.D., Schnabel R.D., Murdoch B.M. *et al.* (2007) Whole genome linkage disequilibrium maps in cattle. *BMC Genetics* **8,** 74–85.

Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157,** 1819–29.

Miller S.A., Dykes D.D. & Polesky H.F. (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research* **16,** 1215.


doi:10.1111/j.1365-2052.2009.02011.x

Nilsen H., Hayes B., Berg P. *et al.* (2008) Construction of a dense SNP map for bovine chromosome 6 to assist the assembly of the bovine genome sequence. *Animal Genetics* **39,** 97–104.

Nordborg M. & Tavaré S. (2002) Linkage disequilibrium: What history has to tell us. *Trends in Genetics* **18,** 83–90.

Odani M., Narita A., Watanabe T. *et al.* (2006) Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. *Animal Genetics* **37,** 1–6.

Patil N., Berno A.J., Hinds D.A. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294,** 1719–23.

Prasad A., Schnabel R.D., McKay S.D. *et al.* (2008) Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. *Animal Genetics* **39,** 597–605.

Pritchard J.K. & Przeworski M. (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69,** 1–14.

Reich D.E. & Lander E.S. (2001) On the allelic spectrum of human disease. *Trends in Genetics* **17,** 502–10.

Reich D.E., Cargill M., Bolk S. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* **411,** 199–204.

Sabeti P.C., Reich D.E., Higgins J.M. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–7.

Sargolzaei M., Schenkel F.S., Jansen G.B. & Schaeffer L.R. (2008) Extent of linkage disequlibrium in Holstein cattle in North America. *Journal of Dairy Science* **91,** 2106–17.

Scheet P. & Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78,** 629–44.

Simianer H., Szyda J., Ramon G. & Lien S. (1997) Evidence for individual and between family variability of the recombination rate. *Mammalian Genome* **8,** 830–5.

Snelling W.M., Chiu R., Schein J.E. *et al.* (2007) A physical map of the bovine genome. *Genome Biology* **8,** R165.

Spelman R.J. & Coppieters W. (2006) *Linkage disequilibrium in the New Zealand Jersey population*. CD-ROM Communication No. 22-21 in Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brazil.

Sorensen A.C., Sorensen M.K. & Berg P. (2005) Inbreeding in Danish dairy cattle breeds. *Journal of Dairy Science* **88,** 1865–72.

Sved J.A. (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2,** 125–41.

Tenesa A., Knott S.A., Ward D. *et al.* (2003) Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *Journal of Animal Science* **81,** 617–23.

Tenesa A., Navarro P., Hayes B.J. *et al.* (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17,** 520–6.

Terwilliger J.D., Haghighi F., Hiekkalinna T.S. & Göring H.H.H. (2002) A biased assessment of the use of SNPs in human complex traits. *Current Opinion in Genetics & Development* **12,** 726–34.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nautre* **437,** 1299–320.

Van Laere A.S., Nguyen M., Braunschweig M. *et al.* (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425,** 832–6.

Villa-Angulo R., Matukumalli L.K., Gill C.A. *et al.* (2009) High-resolution haplotype block structure in the cattle genome. *BMC Genetics* **10,** 19.

Wang N., Akey J.M., Zhang K. *et al.* (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *American Journal of Human Genetics* **71,** 1227–34.

Wright S. (1938) Size of population and breeding structure in relation to evolution. *Science (Wash DC)* **87,** 430–1.

Young C.W. & Seykora A.J. (1996) Estimates of inbreeding and relationship among registered Holstein females in the United States. *Journal of Dairy Science* **79,** 502–5.

Zhang K., Deng M., Chen T., Waterman M.S. & Sun F. (2002) A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 7335–9.

Zhang K., Sun F., Waterman M.S. & Chen T. (2003) Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *American Journal of Human Genetics* **73,** 63–73.


doi:10.1111/j.1365-2052.2009.02011.x