# Extent of linkage disequilibrium in a large cattle population of Western Africa and consequences for association studies.

S. Thevenon, Guiguigbaza-Kossigan Dayo, S. Sylla, Issa Sidibé, Daphné Berthier, Hortense Legros, Didier Boichard, Andre A. Eggen, Mathieu M. Gautier

doi:10.1111/j.1365-2052.2007.01601.x

# The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies

**S. Thévenon*,†,‡, G. K. Dayo*,†,‡, S. Sylla‡, I. Sidibe‡, D. Berthier*,†, H. Legros§, D. Boichard¶, A. Eggen¶ and M. Gautier¶**

*UMR Trypanosomes, CIRAD, Montpellier, F-34398 France. †UMR Trypanosomes, IRD, Montpellier, F-34398 France. ‡URBIO, CIRDES, Bobo-Dioulasso 01, Burkina Faso. §Labogena, Jouy-en-Josas F-78352, France. ¶DGA, INRA, Jouy-en-Josas F-78352, France

**Summary**

Several previous studies concluded that linkage disequilibrium (LD) in livestock populations from developed countries originated from the impact of strong selection. Here, we assessed the extent of LD in a cattle population from western Africa that was bred in an extensive farming system. The analyses were performed on 363 individuals in a *Bos indicus* × *Bos taurus* population using 42 microsatellite markers on BTA04, BTA07 and BTA13. A high level of expected heterozygosity (0.71), a high mean number of alleles per locus (9.7) and a mild shift in Hardy–Weinberg equilibrium were found. Linkage disequilibrium extended over shorter distances than what has been observed in cattle from developed countries. Effective population size was assessed using two methods; both methods produced large values: 1388 when considering heterozygosity (assuming a mutation rate of $10^{-3}$) and 2344 when considering LD on whole linkage groups (assuming a constant population size over generations). However, analysing the decay of LD as a function of marker spacing indicated a decreasing trend in effective population size over generations. This decrease could be explained by increasing selective pressure and/or by an admixture process. Finally, LD extended over small distances, which suggested that whole-genome scans will require a large number of markers. However, association studies using such populations will be effective.

**Keywords** cattle, effective population size, linkage disequilibrium, microsatellites, population history.

## Introduction

Characterizing the extent of marker-marker linkage disequilibrium (LD) provides valuable information when predicting LD between a marker and an allelic variant underlying a trait of interest (QTN) (Kohn *et al.* 2000; Grisart *et al.* 2002; Tanaka *et al.* 2005). Such information is also useful when estimating marker density required in association studies. Additionally, because historical population events influence the pattern of LD among markers, some population characteristics might be inferred from the characterization of the extent of LD (Pritchard & Przeworski 2001; Hayes *et al.* 2003; Abecasis *et al.* 2005; Verardi *et al.* 2006).

In livestock, most studies have focused on populations subjected to intensive breeding typical of developed countries. As expected from their history, these studies have led to the identification of a large amount of LD across the genome. In a pioneering study using a dairy cattle population, Farnir *et al.* (2000) observed a high level of LD for markers located <5 cM apart (average $D'$>0.5, estimated by the multi-allelic Lewontin $D'$ measure, Lewontin (1964)). Similarly, Tenesa *et al.* (2003) estimated an average $D'$ value of 0.44 in another dairy cattle population, while Nsengimana *et al.* (2004) reported a $D'$ value reaching 0.60–0.80 in a commercial pig breed for marker pairs less than 5 cM apart. Although the $D'$ measure is biased upward when considering multiallelic markers, these studies show that LD in livestock extends over larger distances than in human populations for which the extent of LD decreases rapidly below 400 kb Abecasis *et al.* (2001), except in some

Address for correspondence

S. Thévenon, UMR Trypanosomes, TA A-17/G, CIRAD, Campus International de Baillarguet, 34398 Montpellier cedex 5, France.
E-mail: sophie.thevenon@cirad.fr

*Accepted for publication 4 March 2007*

small populations where LD can extend over 5 Mb (Pritchard & Przeworski 2001; Johansson *et al.* 2005). Differences among human and livestock populations can be explained by recent genetic improvement programmes in livestock, whereby intensive selection has resulted in an increase in inbreeding and thus small effective population sizes ($N_e$). For instance, only a few bulls contribute to a large proportion of the French dairy cattle population, resulting in a $N_e$ value much smaller than 100 individuals (Boichard *et al.* 1996).

However, the extent of LD has not yet been assessed for livestock populations bred in extensive systems, but is expected to be smaller due to larger effective population sizes. The aim of this study was to assess the extent of LD in a large mildly selected cattle population from Western Africa under an extensive breeding system. The data consisted of genotypes from 46 microsatellite markers on three bovine chromosomes for 363 individuals assumed to be unrelated.

The animals in the geographic region of interest originated from a cross between a zebu breed (*Bos indicus*) and the Baoule taurine breed (*Bos taurus*, a short-horn cattle breed). A predominance of zebu traits is observed in these animals, including coat colours, hump sizes and horn shapes. The existence of historical and continuous hybridization in Africa is highlighted by several arguments:

**1** West African zebus carry mitochondrial DNA (mtDNA) from taurine origin (reviewed by MacHugh *et al.* 1997).

**2** *B. indicus* Y chromosomes segregate in some taurine populations (MacHugh *et al.* 1997; Hanotte *et al.* 2000).

**3** Zebu alleles at microsatellite markers segregate in taurine populations and vice versa (MacHugh *et al.* 1997; Moazami-Goudarzi *et al.* 2001; Hanotte *et al.* 2002; Freeman *et al.* 2004; Ibeagha-Awemu *et al.* 2004; Freeman *et al.* 2006a).

Reviews by MacHugh *et al.* (1997) and Hanotte *et al.* (2002) indicate that hybridization between *B. indicus*, originating from the Indus Valley and African *B. taurus* must have started around 700 AD with the spread of *B. indicus*. An amplification of this continuous hybridization process may have occurred during the rinder pest epidemic at the end of the 19th century, with taurine individuals being more susceptible to this disease.

## Materials and methods

### Animals

Three hundred and sixty-three cattle were sampled in the south-west of Burkina Faso (longitude: from 4°35′49″ to 4°56′55″W, latitude: from 10°3′39″ to 10°15′16″N) at the border of the tse-tse belt (Hendrickx *et al.* 2004). All animals in this study lived in the same geographic area under a similar agro-ecological context (Soudano-Guinean climate with trypanosomiasis pressure). They belonged to 47 herds (from 1 to 46 individuals sampled per herd), which were bred by Peul pastors under a typical pastoral extensive breeding system with transhumance between the dry and wet seasons. Random mating occurred between males and females.

### Genotyping

Genomic DNA was extracted from blood samples using the Promega Wizard Kit, and DNA was stored at −20 °C. Nineteen, 17 and 10 microsatellite markers distributed over 46 cM on chromosome BTA04, 43 cM on BTA07 and 20 cM on BTA13 were chosen (Table S1) from available genetic map information (Ihara *et al.* 2004). The average marker spacing over the three regions was 2.53 cM (from 0.32 to 8.23) (Ihara *et al.* 2004), which was smaller than average marker spacing in studies by Farnir *et al.* (2000); McRae *et al.* (2002); Tenesa *et al.* (2003) and Nsengimana *et al.* (2004) (13.4, 20, 5.25 and 5.23 cM respectively) and slightly higher than the one in the study by Harmegnies *et al.* (2006) (1.97 cM). Fluorescent multiplex PCR used the PCR Qiagen® kit according to the manufacturer's recommendations. PCR products were then run on a ABI3730x1 Genetic Analyser sequencer (Applied Biosystems). Raw data were analysed with GENEMAPPER v3.7 software (Applied Biosystems).

### Population structure analysis

Allele numbers, allele frequencies and gene diversity were estimated for each locus using GENETIX 4.03 (http://www.genetix.univ-montp2.fr) and GENEPOP 3.4c (Raymond & Rousset 1995). Additionally, observed heterozygosity and gene diversity were estimated both per locus and over all loci (Nei 1987). Departures from Hardy–Weinberg equilibrium over all loci were evaluated using Fisher's method (GENEPOP 3.4c) and permutations (GENETIX 4.03).

The unbiased estimator ($\hat{f}$) of Wright's inbreeding coefficient $F_{IS}$ was calculated according to Weir & Cockerham (1984). The empirical distribution of $F_{IS}$ under the null hypothesis (Hardy–Weinberg equilibrium) was obtained using 10 000 permutations (GENETIX 4.03). While the $F_{IS}$ estimate provides information on the present breeding system, insights on the population admixture history can be inferred from the admixture assessment implemented in STRUCTURE 2.1 (Pritchard *et al.* 2000). STRUCTURE employs a model-based clustering method that utilizes a Monte Carlo Markov Chain to assign individuals to $k$ populations and estimate the posterior distribution of each individual's admixture coefficient. Version 2.1 of STRUCTURE takes into account the correlations that arise between linked markers as the result of admixture or hybridization. Because genotyping information for the putative parental populations was not available, we hypothesized $k$ parental unknown populations ($k$ varying from 1 to 5). When working with

complex data sets, Garnier *et al.* (2004) suggested choosing the *k*-value that maximizes the gain of information as [Ln $P(D)_k$ −Ln $P(D)_{k-1}$], where Ln $P(D)$ represents the posterior probability of the data. STRUCTURE 2.1 was run five times with a burn-in period of $5 \times 10^4$ iterations followed by $10^5$ iterations for each *k*-value, considering both correlated and uncorrelated allelic frequencies between the parental populations.

## Haplotype analysis

*Map order.* Microsatellite sequences were anchored to bovine whole-genome sequence assembly BUILD 3.1 (http://www.hgsc.bcm.tmc.edu/projects/bovine/) using BLAST (Altschul *et al.* 1997). A radiation hybrid (RH) map was built for each linkage group using the 3000-rad RH panel described by Williams *et al.* (2002). RH data were analysed using CARTHAGÈNE (de Givry *et al.* 2005) together with other published vectors (Williams *et al.* 2002) when some discrepancies had to be resolved. Finally, all markers were PCR-screened on the BAC INRA library as described previously (Eggen *et al.* 2001), which made it possible to identify contigs mapping to the region of interest on the INRA BAC-based first-generation bovine physical map published by Schibler *et al.* (2004).

*Haplotype inference.* Because no pedigree data were available, we inferred haplotypes using PHASE 2.1 (Stephens *et al.* 2001; Stephens & Donnelly 2003). Three values for the parameter $\rho = 4N_e c$ (where $N_e$ was the effective population size and $c$ the recombination probability per adjacent base pair) were considered: $\rho = 0.4$, $\rho = 0.04$ and $\rho = 0.004$. Assuming $c = 10^{-8}$ (i.e. 1 cM ≡ 1 Mb; see results) these values corresponded to $N_e = 10\,000$, 1000 and 100 respectively. We also relaxed the assumption of the stepwise mutation model because mutations may be neglected in comparison with recombination events at the considered scale (Shifman & Darvasi 2001).

Fifteen runs were performed per chromosome (five for each of the three $\rho$ values). Expected haplotype frequencies in the theoretical population, calculated for each run, were compared to check the reconstruction consistency among the different runs.

## Linkage disequilibrium estimation

The significance of the genotypic LD between all pairs of markers (both syntenic and non-syntenic) was tested using Fisher's exact test as implemented in GENEPOP 3.4c (Raymond & Rousset 1995).

Gametic pairwise LD measures were derived from the standard measure of LD between two alleles at two different loci: $D_{ij} = p(A_iB_j) - p(A_i)p(B_j)$, where $p(A_i)$ is the frequency of allele $A_i$ at locus $A$, $p(B_j)$ the frequency of allele $B_j$ at locus $B$ and $p(A_iB_j)$ the frequency of haplotype $A_iB_j$ in the population. Three different multiallelic LD measures were then considered in the study.

First, multi-allelic $D'$ was measured as (Lewontin 1964; Hedrick 1987):

$$D' = \sum_{i=1}^{k}\sum_{j=1}^{l} p(A_i)p(B_j)\left|\frac{D_{ij}}{D_{ij}^{\max}}\right|,$$

where $k$ and $l$ were the number of alleles for markers $A$ and $B$ respectively and

$$D_{ij}^{\max} = \min[p(A_i)p(B_j), (1-p(A_i))(1-p(B_j))] \quad \text{when } D_{ij} < 0 \text{ and}$$

$$D_{ij}^{\max} = \min[p(A_i)(1-p(B_j)), p(B_j)(1-p(A_i))] \quad \text{when } D_{ij} \geq 0.$$

Second, multi-allelic $r^2$ was measured as

$$r^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{D_{ij}^2}{(1-p(A_i))(1-p(B_j))}$$

(Hill & Robertson 1968). Finally, multi-allelic $\chi^{2'}$ was measured as

$$\chi^{2\prime} = \frac{\chi^2}{2N(m-1)}$$

(Yamazaki 1977), where $2N$ was the number of haplotypes considered to compute the corresponding pairwise measure, $m = \min(k,l)$ and $\chi^2$ represented the chi-squared statistic defined as

$$\chi^2 = 2N\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{D_{ij}^2}{p(A_i)p(B_j)},$$

(Hill 1975); $\chi^2$ was asymptotically distributed according to a chi-square distribution with $(k-1)(m-1)$ degrees of freedom. The three multiallelic LD measures defined above ($D'$, $r^2$ and $\chi^{2'}$) were computed using POWERMARKER V3.23 (Liu & Muse 2005) for each reconstructed haplotype.

## Model fitting and parameter estimation

All statistical analyses were performed using R software (Ihaka & Gentleman 1996). The decay of LD with genetic distance was fitted with two models.

First, we used the linear model

$$\mathrm{LD}_{ijkl} = \mu + b\ln(2\theta_i) + c_j + r_k + a_l + c_j b\ln(2\theta_i) + r_k b\ln(2\theta_i) + a_l b\ln(2\theta_i) + c_j r_k + c_j a_r + r_k a_l + \varepsilon_{ijkl} \quad (1)$$

(Nsengimana *et al.* 2004), where $\mathrm{LD}_{ijkl}$ was the LD measure value for marker pair $i$ composed of two markers mapping to chromosome $j$ with a recombination rate $\theta_i$ between them ($\theta_i$ was derived from the genetic distance using the Haldane mapping function). Marker phases used to compute LD were estimated from phase during the run $l$ performed with a

corresponding $\rho$ value $k$. $\mu$ was the average LD value, $b$ was the constant associated to $\theta_i$, $c_j$ was the effect of chromosome $j$, $r_k$ was the effect of the $\rho$ value $k$, $a_l$ was the effect of the run $l$ and $\varepsilon_{ijkl}$ was the residual. As the variance of the LD estimation was larger for small genetic distance due to the stochastic process of genes genealogy, the obtained residuals did not fulfil the assumptions of homoscedasticy and normality; thus, we also ran the same models on the transformed ln(LD). Non-significant effects, beginning with the interactions, were progressively removed from the model to simplify it.

Secondly, we considered a non-linear model (Sved 1971) correcting for sample size (Weir & Hill 1980; Harmegnies *et al.* 2006):

$$\mathrm{LD}_{ij} = 1/(1 + 4\beta_j c_i) + \varepsilon_{ij} + 1/2n$$

where $\mathrm{LD}_{ij}$ was the LD measure for marker pair $i$ composed of markers mapping to chromosome $j$ and spaced from a distance $c_i$ between them, $\beta_j$ was the estimation of the parameter $\beta$ (see below) for the chromosome $j$, $n$ was the number of diploid individuals and $\varepsilon_{ij}$ was the residual. $\beta$ was related to the effective population size and, according to Sved (1971) and Zhao *et al.* (2005), $\beta$ was a good estimator of $N_e$ when multi-allelic $r^2$ was used as a measure of LD and assuming a population with a constant effective population size over generations and no substructure. $\beta$ was estimated using a non-linear least-square approximation.

### Effective population size

The effective population size was estimated according to two independent methods. First, we considered the model by Sved (1971), described above, which corresponds to an inbreeding model assuming the population is not undergoing genetic drift or mutation. When considering the whole range of recombination rates, this model additionally assumes a constant $N_e$ over generations. Following Hayes *et al.* (2003), LD value for markers separated by $c$ Morgan reflects $N_e$ $1/2c$ generations ago (at most 50 generations in our experiment because the minimal genetic distance between markers was 0.01 Morgan). We estimated historical $N_e$ by averaging LD by distance classes.

The second model assumed mutation-drift equilibrium and used current marker polymorphism to estimate $N_e$. The composite parameter $\theta_{N_e} = 4N_e\mu$ (where $\mu$ was the mutation rate per locus per generation and $N_e$ was the long-term effective population size) can be estimated from the formulae

$$\theta_{N_e} = \frac{1}{2}\left(\frac{1}{(1 - H_e)^2} - 1\right),$$

where $H_e$ stands for the expected heterozygosity (Ohta & Kimura 1973). Because this estimator of $\theta_{N_e}$ was upwardly biased because of the non-linear transformation of the equation (Wang 2005), we used the unbiased estimator $\theta_F$ proposed by Xu & Fu (2004) for microsatellite markers and defined as:

$$\theta_F = \left(1.1313 + \frac{3.4482}{n} + \frac{28.2878}{n^2}\right)\theta_{N_e} + 0.3998\sqrt{\theta_{N_e}} = \frac{1}{2}\left(\frac{1}{(1 - H_e)^2} - 1\right) \text{ for } \theta_F \leq 15.$$

Xu & Fu (2004) noticed that $N_e$ can be overestimated if mutational events depart from a strict stepwise model. However, the magnitude is not strongly affected and $\theta_F$ has a much smaller sampling variance than the estimator based on variance in allele sizes. No published studies assessed $\mu$ for microsatellite markers in cattle, and mutation rates of $10^{-3}$ and $10^{-4}$ per locus per generation were used (Dallas 1992; Weber & Wong 1993).

## Results

### Genetic variability and population structure

Of the 46 microsatellite markers (Table S1), two (*DIK026* and *DIK2740*) could not be amplified. The 44 other markers were polymorphic, with a mean number of alleles per locus equal to 9.7 (range: 3–23; Table S1). The gene diversity over all 363 individuals and loci was 0.71 (Table S1). Because microsatellites *DIK4290* and *DIK4542* showed a very high $F_{IS}$ (>0.5) compared with other loci and a high genotyping failure rate, we suspected the existence of segregating null alleles. Thus, these markers were discarded and 42 microsatellite markers were used in subsequent analyses. This led to a decrease of the mean $F_{IS}$ from 0.07 to 0.05. However, the overall $F_{IS}$ remained highly significant ($P$-value < 0.001).

To test for population structure and avoid bias due to genetic linkage between loci, we performed complementary analyses using three markers per chromosomes (nine in total, *DIK4682*, *DIK4236*, *MGTG4B*, *DIK2819*, *BMS904*, *DIK630*, *DIK2309*, *DIK537*, *DIK2117*), one at each end of the chosen linkage group and one in the middle, according to their low rate of genotyping failure. At a 5% threshold, the mean $F_{IS}$ was found non-significant using GENETIX software ($F_{IS} = 0.01$, $P$-value = 0.083), and just significant when computing the Fisher's method with GENEPOP ($P$-value = 0.046). Thus, as expected when the breeding system is extensive, the mating system was close to panmixia, with little substructure.

Using genotyping data of the 42 markers and the admixture assessment implemented in STRUCTURE, two or three parental populations seem to be the most likely, with the most consistent gain in information for $k = 2$ (for correlated allele frequencies) and $k = 3$ (for independent allele frequencies) (see Fig. 1). Results obtained assuming correlated allele frequencies showed a large variation among

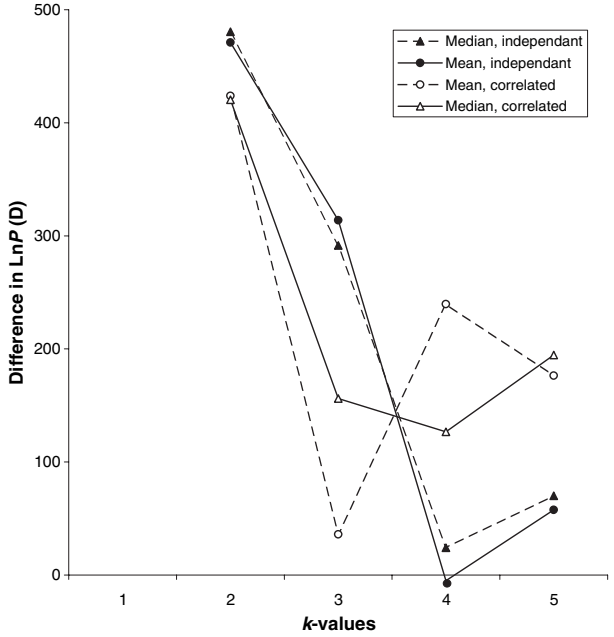


**Figure 1** Differences in the posterior probability of *k* ancestral populations, calculated from the mean and the median of five runs implemented in STRUCTURE, for correlated and independent allele frequencies.

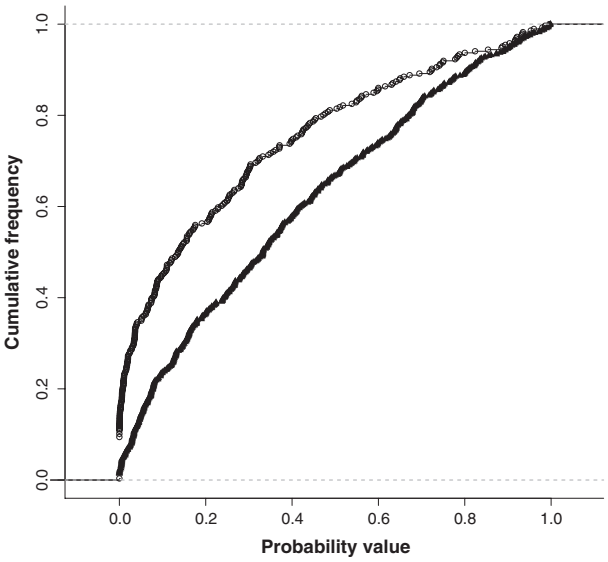


**Figure 2** Cumulative frequency of the *P*-values (GENEPOP); ○: syntenic loci; ▲: non-syntenic loci.

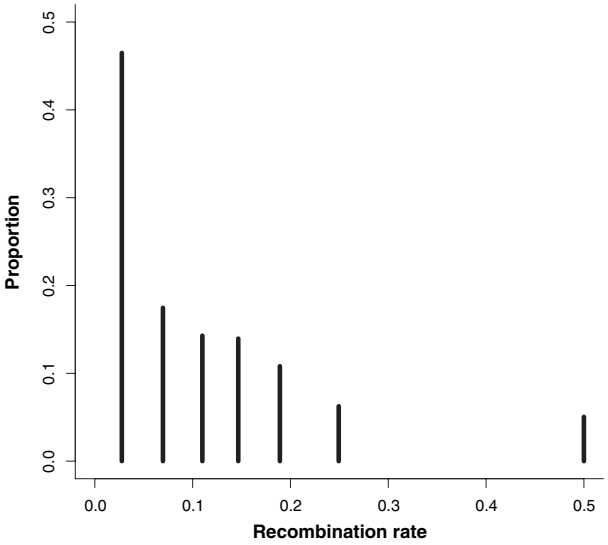


**Figure 3** Proportion of marker pairs with a *P*-value smaller than 0.01 depending on recombination rate.

reconstruction runs as shown by the influence in using the means instead of the medians of the five run values for a given *k* number of parental populations. In any case, results from the different runs indicated that individuals from our study were historically admixed with a genome probably originating equally from the two to three parental populations. Nevertheless, the population can be considered currently as homogenous, with a mating system close to panmixia.

## Linkage disequilibrium

Cumulative frequencies of the *P*-values obtained with GENEPOP are shown in Fig. 2 for the 286 syntenic pairs (markers on the same chromosome, 105 pairs on BTA04, 136 on BTA07 and 45 on BTA13) and the 575 non-syntenic pairs (markers on different chromosomes). As expected, significant LD was observed more frequently for syntenic markers than for non-syntenic markers. The proportion of marker pairs with a *P*-value < 0.01 depending on the recombination rate is shown in Fig. 3. Clearly, LD was significant for a large proportion of marker pairs when they were < 5 cM apart. However, the proportion of significant LD for non-syntenic pairs was slightly larger than expected by chance.

*Map construction and order checking.* As shown in Table S1, the order of markers in the RH map was similar to that in the previous linkage map, except for inversions concerning two pairs of closely linked markers on BTA07: *CSSM057* and *DIK2256* and *MNB15* and *DIK2407*. Analyses on the Btau3.1 bovine genome assembly (Table S1) confirmed the first inversion, while the second one could not be resolved because the location of *MNB15* was not found on the assembly. Conversely, a group of three markers mapping to BTA13 were inverted in the Btau3.1 assembly compared with their order on the linkage and RH maps. These two discrepancies on BTA07 and BTA13 were resolved by BAC-screening results, which confirmed the RH panel order. Thus, we used the RH map order in our analyses. A simple linear extrapolation was done to correct the distance for the

two inverted pairs of markers. The ratio $b$ between the marker distance in cM and in Mb (as estimated from the Btau3.1 bovine genome sequence assembly) was roughly equal to 1, with $\hat{b} = 1.08$ (SD = 0.17) for BTA04, $\hat{b} = 0.78$ (SD = 0.06) for BTA07 and $\hat{b} = 1.06$ (SD = 0.16) for BTA13.

*Haplotype inference.* Results from the different phase runs were consistent, with the same estimated population haplotypes at high frequencies. More precisely, we obtained exactly the same haplotyping results for 54, 34 and 18% of the individuals across the 15 runs for BTA13, BTA07 and BTA04 linkage groups respectively and 26, 17 and 13% of the animals had the most likely haplotype pair probability above 0.9 for BTA13, BTA07 and BTA04 respectively. As expected, phase performed better for the BTA13 linkage group, with fewer markers across a smaller genetic distance than BTA07 and BT04.

*Linkage disequilibrium measures.* As shown in Table 1, mean $D'$ was 0.255 (SD = 0.095) for markers <3 cM apart and dropped to 0.158 above 25 cM (SD = 0.053). For markers < 3 cM apart, average $r^2$ and $\chi^{2'}$ were 0.010 (SD = 0.007) and 0.060 (SD = 0.039) respectively. Interestingly, all measures computed from haplotype data were highly correlated ($P$-value < 0.001) and were also significantly ($P$-value < 0.001) correlated to the genotypic $P$-value computed in GENEPOP.

*Model fitting.* No significant effect on LD was found for the $\rho$ value, the phase run and their interactions with other parameters in the model. The mean and the standard deviation of the measure $\chi^{2'}$ for the three linkage groups are shown in Fig. 4. For all measures, including the $\chi^2$ $P$-value, effects of recombination rate and linkage group were highly significant ($P$-value < $10^{-15}$ and $P$-value < 0.001 respectively), and they explained 35% of the total variance for $r^2$, 19% for $D'$ and 19% for $\chi^{2'}$.

We then fitted the non-linear model described in Materials and methods, and found $\hat{\beta} = 2344.3$ (SD = 118.7) over the three chromosomes, and, more particularly, $\hat{\beta}_{BTA04} = 2{,}201.1$ (SD = 206.9) when considering BTA04 only, $\hat{\beta}_{BTA07} = 2{,}423.1$ (SD = 184.3) for BTA07 and $\hat{\beta}_{BTA13} = 2{,}240.7$ (SD = 234.6) for BTA13. A significant effect of linkage group on $\hat{\beta}$ ($P$-value < $10^{-5}$) was found.

### $N_e$ estimation

We estimated $\hat{\beta} = N_e = 2344$ (SD = 119) under the non-linear model describing LD evolution along whole linkage groups. On the other hand, estimates of $N_e$ from the expected heterozygosity for 9 and 42 loci respectively (see above) were 18 925 and 13 877 individuals assuming a mutation rate of $10^{-4}$ per locus per generation, and 1893 and 1388 individuals assuming a mutation rate of $10^{-3}$. Moreover, the evolution of $N_e$ over time was consistent with a decrease in population size over generations (Fig. 5).

## Discussion

In this study, significant LD was observed for pairs of linked markers across three linkage groups in a west-African cattle population bred under extensive conditions, when considering both genotype and haplotype data. There was no indication of bias introduced by the haplotype reconstruction using PHASE software, as the reconstruction of haplotypes was consistent across the five independent runs performed for each chromosome and each $\rho$-value. $P$-values for LD estimated from haplotype and genotype data were significantly correlated. An excess of heterozygotes might introduce an important bias in haplotype reconstruction. As no strong departures from Hardy–Weinberg equilibrium were observed for the markers used in our analysis (see below), we excluded this potential pitfall.

As expected (Zhao *et al.* 2005), $\chi^{2'}$, which was proposed as the best predictor of the useful LD for microsatellite markers, had intermediate values between $D'$ and $r^2$ estimates. $D'$ and $r^2$ were upwardly and downwardly biased respectively when considering multi-allelic loci (McRae *et al.* 2002; Tenesa *et al.* 2003; Heifetz *et al.* 2005; Zhao *et al.* 2005). In particular, $D'$ was inflated with rare alleles (Ardlie *et al.* 2002; McRae *et al.* 2002).

Linkage disequilibrium decreased when genetic distance increased. Nevertheless, linear and non-linear models

**Table 1** Mean and standard deviation (SD) of linkage disequilibrium measures estimated from haplotype data for different distance ranges.

| Distance | | $r^2$ | | $D'$ | | $\chi^{2'}$ | |
|---|---|---|---|---|---|---|---|
| Range (cM) | Mean | Mean | SD | Mean | SD | Mean | SD |
| (0–3) | 1.7 | 0.010 | 0.007 | 0.255 | 0.095 | 0.060 | 0.039 |
| (3–6) | 4.5 | 0.005 | 0.003 | 0.230 | 0.083 | 0.053 | 0.039 |
| (6–9) | 7.6 | 0.005 | 0.002 | 0.224 | 0.093 | 0.040 | 0.026 |
| (9–12) | 10.5 | 0.004 | 0.002 | 0.181 | 0.051 | 0.034 | 0.019 |
| (12–15) | 13.4 | 0.003 | 0.001 | 0.177 | 0.056 | 0.035 | 0.019 |
| (15–19) | 16.8 | 0.003 | 0.001 | 0.168 | 0.049 | 0.031 | 0.019 |
| (19–25) | 21.4 | 0.003 | 0.001 | 0.166 | 0.052 | 0.028 | 0.019 |
| (25–45) | 31.3 | 0.002 | 0.001 | 0.158 | 0.053 | 0.024 | 0.013 |

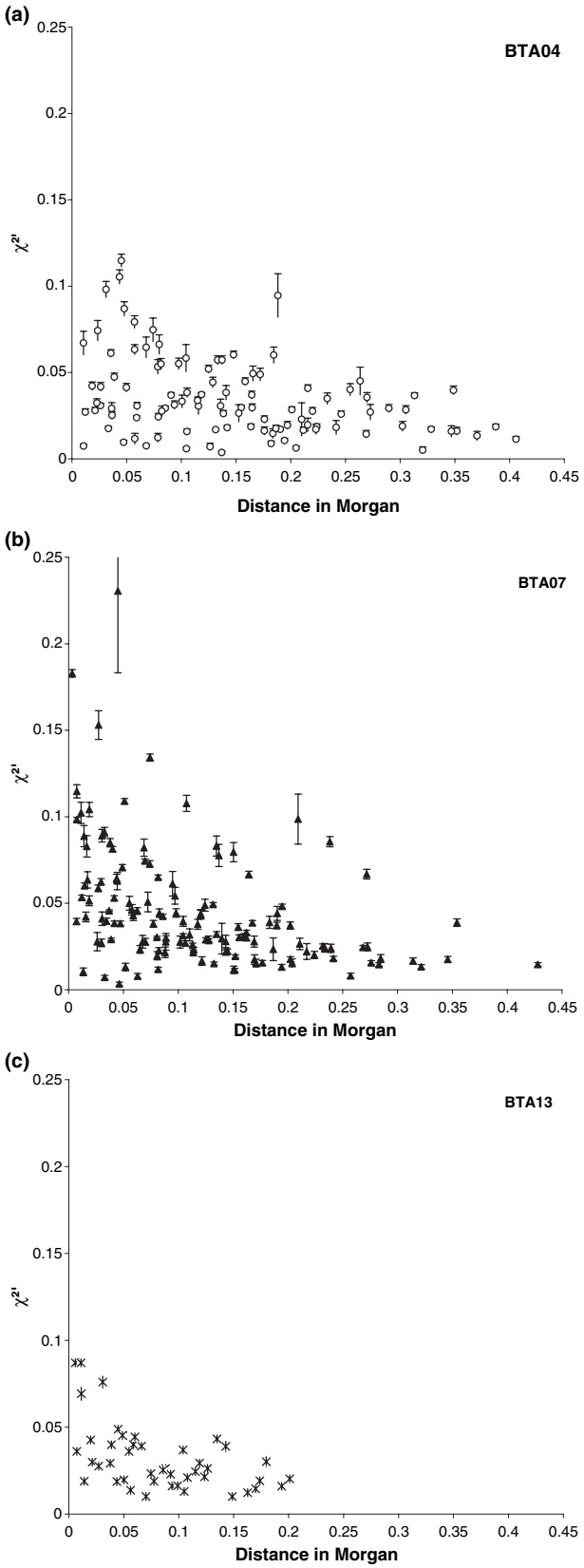


**Figure 4** $\chi^{2'}$ mean and standard deviation calculated over the 15 runs for BTA04 (a), BTA07 (b) and BTA13 (c). The standard deviation is represented by the vertical line (± $\sigma$).

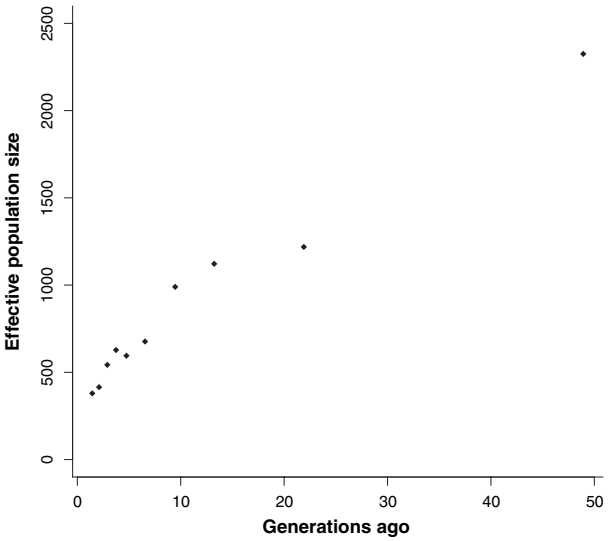


**Figure 5** Estimates of effective population size over time (number of generations).

showed a significant effect of linkage group on the pattern of LD. Looking at the $r^2$ estimator, $\hat{\beta}$ was significantly larger for BTA07 than for BTA04 and BTA13. This difference could be due to a selection process acting on the two latter chromosomal regions (Pritchard & Przeworski 2001), due to human manipulations or environmental conditions. Hanotte *et al.* (2003) identified QTL underlying tolerance to trypanosomiasis in an experimental $F_2$ population of Boran zebu (*B. indicus*) and N'dama taurine (*B. taurus*). Trypanosomiasis is the main parasitic constraint on livestock breeding in sub-Saharan Africa (Swallow 2000). In particular, a highly significant QTL was found to map to BTA04.

The estimated LD level was much smaller in our population than in other livestock populations (Farnir *et al.* 2000; McRae *et al.* 2002; Tenesa *et al.* 2003; Nsengimana *et al.* 2004; Harmegnies *et al.* 2006). This highlights the strong disparity between highly selected cattle populations from developed countries and large poorly selected populations from developing countries. The pattern of LD in other studies shows the impact of strong selection (Farnir *et al.* 2000; Nsengimana *et al.* 2004; Harmegnies *et al.* 2006). Compared with human populations, the pattern of LD in highly selected livestock population tends to be higher at large distances and relatively smaller at short distances, which is characteristic of a decreasing population size (Hayes *et al.* 2003). In our data set, LD for closely linked markers was difficult to infer because of a low marker density. In any case, the pattern of LD in our population suggests small genetic drift associated with a small decrease in population size over generations, assuming our population remained homogeneous during this period (see below).

A mild shift from Hardy–Weinberg equilibrium was found, which could explain the LD between unlinked markers that was found significantly more frequently than

expected by chance. However, the $F_{IS}$ value was small and no population structure was detected. This population was currently homogeneous, although it originated probably from hybridization between two breeds. It was considered as a unique population in the analyses.

Hybridization was assumed after the animals' phenotypes and farmers' declarations and is further supported by the estimated LD and expected heterozygosity. In *B. taurus* or *B. indicus* cattle populations from western Africa, expected heterozygosity ranged from 0.46 to 0.54 in Baoule and to around 0.61 in Peul Fulani animals from the Burkina Faso and Borgou breeds (a cross-bred between *B. indicus* and *B. taurus*) (Freeman *et al.* 2004, 2006a). In our study, the expected heterozygosity was larger and close to that estimated in cattle populations in Turkey and in the northern part of the Arabian Peninsula (Freeman *et al.* 2006a), and in admixed cattle populations from Cameroun and Nigeria (Ibeagha-Awemu *et al.* 2004). The mean number of alleles per locus was also high in our study. An accurate estimation of $N_e$ from heterozygozity relies on several assumptions (reviewed in Wang 2005): assessment of long-term $N_e$, single-step mutation model of neutral alleles, a population without substructure, mutation-drift equilibrium and the ability to assess the mutation rate $\mu$. Our high $N_e$ estimation may be related to a large effective population size and/or be the result of hybridization. Hybridization has already been hypothesized to explain the high level of diversity found in cattle breeds at the border of the tse-tse belt (Freeman *et al.* 2004). The estimation from LD assumes an isolated population at equilibrium with a constant $N_e$. However, when change in population size occurs, the average distance between marker pairs provides estimates of $N_e$ at different times because the closer the loci, the longer the time required for LD to disappear (Hill 1981; Hayes *et al.* 2003). In our data set, the estimation of $N_e$ from LD at different distance ranges decreased over generations: this could be explained by a strong selective pressure and/or by an admixture process that results in an increased LD. Increasing selection is likely to occur because of trypanosomiasis pressure in the area, a disease that can decimate sensitive cattle populations (*B. indicus*) (Stewart 1951; Murray *et al.* 1984; Roelants *et al.* 1987). Indeed, Peul pastors have been trying to penetrate wet areas with a high trypanosomiasis prevalence at an increasing rate to find pastures (Lhoste 1991). However, the use of trypanocide drugs is widespread and decreases the selection strength. In addition, there is continuous admixture between *B. taurus* and *B. indicus* (MacHugh *et al.* 1997; Hanotte *et al.* 2002; Freeman *et al.* 2004) that could also explain a decrease in $N_e$ estimated from LD. The proportion of marker pairs in significant LD was close to that observed in cattle populations called West African Hybrid by Freeman *et al.* (2006b); however, the LD decay from our data cannot be compared with this study because Freeman *et al.* (2006b) did not report the decay of LD with genetic distance.

We thus give preference to a stronger influence of admixture on the observed pattern of LD, the large observed heterozygosity and the mean number of alleles. Nevertheless, additional samples from other zebu from Africa, pure zebu from India and taurine individuals and genotyping of closer linked markers would allow a better understanding of the population history.

Characterization of the extent of marker-marker LD provides insights to assess the power of association studies aiming at mapping loci underlying traits of interest. $\chi^{2'}$ measures have been proposed as the best predictor of useful LD when microsatellite markers are used (Zhao *et al.* 2005). They have the same expectation as the proportion of QTL variance that would be explained by one of the markers if the QTL and the marker were at the same distance. In our study, the average value of $\chi^{2'}$ was 0.060 for markers roughly 1.7 cM apart. Abecasis *et al.* (2001) and Pritchard & Przeworski (2001) proposed a criterion for useful LD corresponding to $r^2 \geq 0.1$, meaning that the sample size necessary to detect association at the marker should not exceed 10 times the size of the sample genotyped for the susceptible locus itself. Some but few $\chi^{2'}$ values reached this threshold for closely linked markers. Assuming that a genome scan identifies a candidate genomic region, extensive livestock populations can be useful for fine mapping because the range of LD is smaller than in selected livestock populations. In a highly selected population with a small population size, identification of QTN is complicated by the fact that large chromosome segments, identifiable by a long unique marker haplotype surrounding the favourable allele of the QTN, are quickly swept to high frequency in the population, limiting mapping resolution. In extensive livestock populations, such as the western African population investigated here, the length of a unique marker haplotype surrounding the favourable QTN allele is more limited due to weaker selection and increased probability of recombination. However, the marker density should be chosen preferably with intervals smaller than 1 cM to achieve sufficient power. This should be possible with the development of high-throughput SNP genotyping techniques.

## Acknowledgements

The authors thank the farmers and the CIRDES team for providing the samples. We thank Hélène Hayes (INRA) for reading and correcting the text. We also thank the INRA Department of Animal Genetics, the Laboratoire de Génétique Biochimique et de Cytogénétique (E.P. Cribiu director) and Labogena. G. Dayo was supported by a PhD scholarship from IRD (France). The field study and genotyping were funded by the CORUS programme (Coopération pour la Recherche Universitaire et Scientifique) from French Ministry of Foreign Affaires and by the BRG Programme (Bureau des Ressources Génétiques, France).

We thank the anonymous reviewers and the editor whose relevant comments have helped us to improve the manuscript.

## References


Abecasis G.R., Noguchi E., Heinzmann A. *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *American Journal of Human Genetics* **68,** 191–7.

Abecasis G.R., Ghosh D. & Nichols T.E. (2005) Linkage disequilibrium: ancient history drives the new genetics. *Human Heredity* **59,** 118–24.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25,** 3389–402.

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

Boichard D., Maignel L. & Verrier E. (1996) Analyse généalogique des races bovines laitières françaises. *INRA, Production Animales* **9,** 323–35.

Dallas J.F. (1992) Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. *Mammalian Genome* **3,** 452–6.

Eggen A., Gautier M., Billaut A. *et al.* (2001) Construction and characterization of a bovine BAC library with four genome-equivalent coverage. *Genetics Selection Evolution* **33,** 543–8.

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

Freeman A.R., Meghen C.M., Machugh D.E., Loftus R.T., Achukwi M.D., Bado A., Sauveroche B. & Bradley D.G. (2004) Admixture and diversity in West African cattle populations. *Molecular Ecology* **13,** 3477–87.

Freeman A.R., Bradley D.G., Nagda S., Gibson J.P. & Hanotte O. (2006a) Combination of multiple microsatellite data sets to investigate genetic diversity and admixture of domestic cattle. *Animal Genetics* **37,** 1–9.

Freeman A.R., Hoggart C.J., Hanotte O. & Bradley D.G. (2006b) Assessing the relative ages of admixture in the bovine hybrid zones of Africa and the Near East using X chromosome haplotype mosaicism. *Genetics* **173,** 1503–10.

Garnier S., Alibert P., Audiot P., Prieur B. & Rasplus J.Y. (2004) Isolation by distance and sharp discontinuities in gene frequencies: implications for the phylogeography of an alpine insect species, *Carabus solieri*. *Molecular Ecology* **13,** 1883–97.

de Givry S., Bouchez M., Chabrier P., Milan D. & Schiex T. (2005) CARHTA GENE: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics* **21,** 1703–4.

Grisart B., Coppieters W., Farnir F. *et al.* (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. *Genome Research* **12,** 222–31.

Hanotte O., Tawah C.L., Bradley D.G., Okomo M., Verjee Y., Ochieng J. & Rege J.E. (2000) Geographic distribution and frequency of a taurine *Bos taurus* and an indicine *Bos indicus* Y specific allele amongst sub-saharan African cattle breeds. *Molecular Ecology* **9,** 387–96.

Hanotte O., Bradley D.G., Ochieng J.W., Verjee Y., Hill E.W. & Rege J.E. (2002) African pastoralism: genetic imprints of origins and migrations. *Science* **296,** 336–9.

Hanotte O., Ronin Y., Agaba M. *et al.* (2003) Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 7443–8.

Harmegnies N., Farnir F., Davin F., Buys N., Georges M. & Coppieters W. (2006) Measuring the extent of linkage disequilibrium in commercial pig populations. *Animal Genetics* **37,** 225–31.

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.

Hedrick P.W. (1987) Gametic disequilibrium measures: proceed with caution. *Genetics* **117,** 331–41.

Heifetz E.M., Fulton J.E., O'Sullivan N., Zhao H., Dekkers J.C. & Soller M. (2005) Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics* **171,** 1173–81.

Hendrickx G., De La Rocque S. & Mattioli R.C. (2004) *Long-Term Tse-Tse and Trypanosomiasis Management Options in West Africa*. FAO, Rome.

Hill W.G. (1975) Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8,** 117–26.

Hill W.G. (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* **38,** 209–16.

Hill W.G. & Robertson A. (1968) The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60,** 615–28.

Ibeagha-Awemu E.M., Jann O.C., Weimann C. & Erhardt G. (2004) Genetic diversity, introgression and relationships among West/Central African cattle breeds. *Genetics Selection Evolution* **36,** 673–9.

Ihaka R. & Gentleman R. (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5,** 299–314.

Ihara N., Takasuga A., Mizoshita K. *et al.* (2004) A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Research* **14,** 1987–98.

Johansson A., Vavruch-Nilsson V., Edin-Liljegren A., Sjolander P. & Gyllensten U. (2005) Linkage disequilibrium between microsatellite markers in the Swedish Sami relative to a worldwide selection of populations. *Human Genetics* **116,** 105–13.

Kohn M.H., Pelz H.J. & Wayne R.K. (2000) Natural selection mapping of the warfarin-resistance gene. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 7911–5.

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

Lhoste P. (1991) Cattle genetic resources of West Africa. In: *Cattle Genetic Resources* (Ed. by C.G. Hickman ), pp. 73–89. Elsevier, Amsterdam.

Liu K. & Muse S.V. (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21,** 2128–9.

MacHugh D.E., Shriver M.D., Loftus R.T., Cunningham P. & Bradley D.G. (1997) Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* **146,** 1071–86.

McRae A.F., McEwan J.C., Dodds K.G., Wilson T., Crawford A.M. & Slate J. (2002) Linkage disequilibrium in domestic sheep. *Genetics* **160,** 1113–22.

Moazami-Goudarzi K., Belemsaga D., Ceriotti G. *et al.* (2001) Caractérisation génétique de la race bovine Somba à l'aide de marqueurs moléculaires. *Revue d'Elevage et de Médecine Vétérinaire des Pays Tropicaux* **54,** 129–38.

Murray M., Trail J.C., Davis C.E. & Black S.J. (1984) Genetic resistance to African Trypanosomiasis. *The Journal of Infectious Diseases* **149,** 311–9.

Nei M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY.

Nsengimana J., Baret P., Haley C.S. & Visscher P.M. (2004) Linkage disequilibrium in the domesticated pig. *Genetics* **166,** 1395–404.

Ohta T. & Kimura M. (1973) A model of mutation appropriate to estimate the number of electrophoretic detectable alleles in a finite population. *Genetical Research* **22,** 201–4.

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

Pritchard J.K., Stephens M. & Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–59.

Raymond M. & Rousset F. (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86,** 248–9.

Roelants G.E., Fumoux F., Pinder M., Queval R., Bassinga A. & Authie E. (1987) Identification and selection of cattle naturally resistant to African trypanosomiasis. *Acta Tropica* **44,** 55–66.

Schibler L., Roig A., Mahe M.F., Save J.C., Gautier M., Taourit S., Boichard D., Eggen A. & Cribiu E.P. (2004) A first generation bovine BAC-based physical map. *Genetic Selection Evolution* **36,** 105–22.

Shifman S. & Darvasi A. (2001) The value of isolated populations. *Nature Genetics* **28,** 309–10.

Stephens M. & Donnelly P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* **73,** 1162–9.

Stephens M., Smith N.J. & Donnelly P. (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68,** 978–89.

Stewart J. (1951) The West African Shorthorn cattle. Their value to Africa as trypanosomiasis-resistant animals. *Veterinary Record* **63,** 454–7.

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

Swallow B.M. (2000) *Impacts of Trypanosomiasis on African Agriculture*. FAO (Food and Agricultural Organization of the United Nations), Rome, Italy, pp. 52.

Tanaka G., Matsushita I., Ohashi J. *et al.* (2005) Evaluation of microsatellite markers in association studies: a search for an immune-related susceptibility gene in sarcoidosis. *Immunogenetics* **56,** 861–70.

Tenesa A., Knott S.A., Ward D., Smith D., Williams J.L. & Visscher P.M. (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.

Verardi A., Lucchini V. & Randi E. (2006) Detecting introgressive hybridization between free-ranging domestic dogs and wild wolves (*Canis lupus*) by admixture linkage disequilibrium analysis. *Molecular Ecology* **15,** 2845–55.

Wang J. (2005) Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **360,** 1395–409.

Weber J.L. & Wong C. (1993) Mutation of human short tandem repeats. *Human Molecular Genetics* **2,** 1123–8.

Weir B.S. & Cockerham C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38,** 1358–70.

Weir B.S. & Hill W.G. (1980) Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95,** 477–88.

Williams J.L., Eggen A., Ferretti L. *et al.* (2002) A bovine whole-genome radiation hybrid panel and outline map. *Mammalian Genome* **13,** 469–74.

Xu H. & Fu Y.X. (2004) Estimating effective population size or mutation rate with microsatellites. *Genetics* **166,** 555–63.

Yamazaki T. (1977) The effects of overdominance of linkage in a multilocus system. *Genetics* **86,** 227–36.

Zhao H., Nettleton D., Soller M. & Dekkers J.C. (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genetical Research* **86,** 77–87.

## Supplementary Material

The following supplementary material is available for this article online from http://www.blackwell-synergy.com/doi/full/10.1111/j.1365-2052.2007.01601.x

**Table S1** Summary information on marker positions based on a bovine linkage map, BLAST against the bovine whole-genome sequence (version Btau3.1) and a radiation hybrid map constructed within the study.

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors.