

Linkage Disequilibrium on the Bovine X Chromosome: Characterization and Use in Quantitative Trait Locus Mapping

Cynthia Sandor,* Frédéric Farnir,* Sarah Hansoul,* Wouter Coppieters,* Théo Meuwissen[†] and Michel Georges*¹

*Unit of Animal Genomics, Department of Animal Production, Faculty of Veterinary Medicine and CBIG, University of Liège, 4000 Liège, Belgium and [†]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences (UMB), 1432 Aas, Norway

Manuscript received April 10, 2006
Accepted for publication April 20, 2006

ABSTRACT

We herein demonstrate that in the Holstein–Friesian dairy cattle population, microsatellites are as polymorphic on the X chromosome as on the autosomes but that the level of linkage disequilibrium between these markers is higher on the X chromosome than on the autosomes. The latter observation is not compatible with the small male-to-female ratio that prevails in this population and results in a higher gonosomal than autosomal effective population size. It suggests that the X chromosome undergoes distinct selective or mutational forces. We describe and characterize a novel Markovian approach to exploit this linkage disequilibrium to compute the probability that two chromosomes are identical-by-descent conditional on flanking marker data. We use the ensuing probabilities in a restricted maximum-likelihood approach to search for quantitative trait loci (QTL) affecting 48 traits of importance to the dairy industry and provide evidence for the presence of QTL affecting 5 of these traits on the bovine X chromosome.

EXTENSIVE use of artificial insemination (AI) in mammalian livestock species, especially cattle and pigs, leads to the frequent occurrence of large paternal half-sibships. This pedigree structure has been abundantly exploited to map quantitative trait loci (QTL) influencing a variety of agronomically important phenotypes. Detection of significant differences between the phenotypic means of half-sibs sorted according to the homolog inherited from their sire points toward linked QTL (*e.g.*, GEORGES *et al.* 1995). However, as sires are hemizygous for the X chromosome, this strategy has precluded exploration of the sex chromosomes except for the pseudoautosomal region (*e.g.*, HIENDLER *et al.* 2003; KUHN *et al.* 2003).

It has recently been shown that linkage disequilibrium (LD) extends over unusually long distances in livestock species due to reduced effective population size (N_e) (*e.g.*, FARNIR *et al.* 2000). As a result, significant LD can be detected using the medium-density marker maps available in these species. This can be exploited to increase the power and resolution of QTL mapping (*e.g.*, MEUWISSEN and GODDARD 2000; FARNIR *et al.* 2002). It also opens possibilities to explore correlation between phenotype and marker genotype on the X chromosome, *i.e.*, to map QTL on that chromosome.

In this article, we have (i) quantified the level of microsatellite polymorphism and LD on the X chromo-

some in Holstein–Friesian dairy cattle, (ii) used an approach based on Markov chains to compute identity-by-descent (IBD) probabilities of a pair of X chromosomes conditional on flanking marker data, and (iii) used corresponding IBD probabilities to map QTL on the X chromosome.

MATERIALS AND METHODS

Pedigree material and phenotypes: We used a previously described Holstein–Friesian granddaughter design (GDD) (COPPIETERS *et al.* 1998) comprising 22 paternal half-brother families for a total of 929 bulls. The genealogies of these animals were obtained from Holland Genetics (Arnhem, The Netherlands). The average number of recorded ancestors per bull–dam was 40.4, whereas up to 11 generations separated the bulls from their most distant ancestor. On the basis of available pedigree data and assuming that the founders were unrelated, we estimated the average inbreeding coefficient, F , of the bull–dams at 1.3% (range: 0–14%), and their average kinship coefficient, f , at 4% (range: 0–57%) (FARNIR *et al.* 2000).

For each of these bulls we obtained estimated breeding values (BV) from Holland Genetics (Arnhem, The Netherlands) for 48 phenotypes of interest to the dairy industry. The analyzed phenotypes can be grouped in (i) 5 milk yield and milk composition traits, (ii) 13 conformation traits, (iii) 8 birth and fertility traits, (iv) 10 udder health traits, (v) 2 workability traits, and (vi) 10 productivity traits (Table 1). A detailed description of each of these traits can be found at <http://www.nrs.nl/index-eng.htm>.

Marker genotyping: Using standard procedures (*e.g.*, COPPIETERS *et al.* 1998), the entire GDD was genotyped for 22 X-specific and three pseudoautosomal microsatellite markers, which have all been previously described (SONSTEGARD *et al.* 2001). Marker order and recombination rates between

¹Corresponding author: Unit of Animal Genomics, Faculty of Veterinary Medicine and CBIG, University of Liège (B43), 20 Blvd. de Colonster, 4000 Liège, Belgium. E-mail: michel.georges@ulg.ac.be

TABLE 1
Analyzed phenotypes

Milk (5)	Conformation (13)	Birth/fertility (8)	Udder health (10)	Workability (2)	Productivity (10)
Milk yield (kg)	Angularity	Birth weight	Cell counts	Milking speed	Direct durability
Fat yield (kg)	Body capacity	Calving ease	Fore udder attachment	Temperament	Direct viability cows
Protein yield (kg)	Chest width	Calving ease of the daughters	Front teat placement		Durability
Fat %	Condition score	Fertility index	Rear teat placement		Durable prestation sum
Protein %	Feet and legs	Gestation length	Rear udder height		Index indirect viability
	Foot diagonal	Interval calving—first insemination	Suspensory ligament		Index direct viability
	Frame	Nonreturn rate at 56 days in the daughters	Teat length		Indirect viability cows
	Rear leg set	Stillbirth	Udder		Indirect viability heifers
	Rear leg rear view		Udder depth		Net profit index
	Rump angle		Udder health index		milk production
	Rump width				“Total”
	Stature				
	Weight				

The phenotypes other than milk yield and composition are expressed using normalized scores with mean of 100 and variance of 16 (see <http://www.nrs.nl/index-eng.htm>).

adjacent markers were taken from SONSTEGARD *et al.* (2001) and are shown in Figure 1. We also used phased genotypes available on the same GDD for 202 autosomal microsatellites (FARNIR *et al.* 2000). Sex-averaged recombination rates between autosomal markers were obtained from IHARA *et al.* (2004).

Measuring linkage disequilibrium: Pairwise LD was measured using r^2 as previously described (GRISART *et al.* 2004). Briefly, r^2 was computed as

$$\sum_{i=1}^u \sum_{j=1}^v x_{ij} \frac{(x_{ij} - p_i q_j)^2}{p_i(1 - p_i)q_j(1 - q_j)}, \quad (1)$$

where u and v are the respective numbers of alleles at the two marker loci A and B , p_i and q_j are the population frequencies of alleles A_i and B_j , respectively, and x_{ij} is the observed frequency of haplotype $A_i B_j$.

Computing IBD probabilities conditional on marker genotype for pairs of X chromosomes: The probability of IBD of two X chromosomes at a given map position conditional on multiple linked marker genotypes, Φ_p , was computed according to MEUWISSEN and GODDARD (2001). This requires the computation of the probability of coalescence without recombination, p_C^T , of a chromosome segment of size θ in a population with constant male and female population sizes of, respectively, N_m and N_f , within T generations, where T is the number of generations separating the present from the unrelated base population. This probability is computed using hidden Markov theory, knowing that

$$\mathbf{P}^{t+1} = \mathbf{P}^t \mathbf{M}, \quad (2)$$

where $\mathbf{P}^t = (p_{MM}^t, p_{MF}^t, p_{FF}^t, p_C^t)$, p_{MM}^t is the probability that the two X chromosome segments are two different male chromosomes t generations before present, p_{MF}^t is the probability that the two X chromosomes are one a male the other a female chromosome t generations before present, p_{FF}^t is the probability that the two X chromosomes are two different female chromosomes t generations before present, and p_C^t is the probability that the two X chromosomes have coalesced

between the present and t generations back. $\mathbf{M}$ is a matrix of transition probabilities:

$$\mathbf{M} = \begin{pmatrix} p_{MM \rightarrow MM} & p_{MM \rightarrow MF} & p_{MM \rightarrow FF} & p_{MM \rightarrow C} \\ p_{MF \rightarrow MM} & p_{MF \rightarrow MF} & p_{MF \rightarrow FF} & p_{MF \rightarrow C} \\ p_{FF \rightarrow MM} & p_{FF \rightarrow MF} & p_{FF \rightarrow FF} & p_{FF \rightarrow C} \\ p_{C \rightarrow MM} & p_{C \rightarrow MF} & p_{C \rightarrow FF} & p_{C \rightarrow C} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2}(1 - \theta_f)(1 - \theta_m) & \frac{1}{4}(1 - \frac{1}{N_m})(1 - \theta_m)^2 & \frac{1}{2}(1 - \theta_m)(1 - \theta_f) \\ \frac{1}{4}(1 - \frac{1}{N_m})(1 - \theta_m)^2 & \frac{1}{2}(1 - \theta_m)(1 - \theta_f) & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} (1 - \frac{1}{2N_f})(1 - \theta_f)^2 & \frac{1}{2N_f}(1 - \theta_f)^2 \\ \frac{1}{2}(1 - \frac{1}{2N_f})(1 - \theta_f)^2 & \frac{1}{4N_f}(1 - \theta_f)^2 \\ \frac{1}{4}(1 - \frac{1}{2N_f})(1 - \theta_f)^2 & \frac{1}{4}(\frac{1}{N_m}(1 - \theta_m)^2 + \frac{1}{2N_f}(1 - \theta_f)^2) \\ 0 & 1 \end{pmatrix}. \quad (3)$$

$p_{MM \rightarrow MM}$ is the probability for the two X chromosomes to be different male chromosomes t generations ago and different male chromosomes $t + 1$ generations ago; $p_{MM \rightarrow MF}$ is the probability for the two X chromosomes to be different male chromosomes t generations ago and one a male chromosome and the other a female chromosome $t + 1$ generations ago, etc. . . .; and θ_m and θ_f are male and female recombination rates, respectively. For all but the pseudoautosomal region of the sex chromosomes $\theta_m = 0$. As we are using sires in a GDD, $\mathbf{P}^0$ was set at $(1, 0, 0, 0)$. T was set at 100, N_m at 25, and N_f at 1,000,000, yielding an effective population size N_e of ≈ 100 .

Note that the same approach can also be used to compute coalescence probabilities for autosomal chromosome segments. $\mathbf{M}$ then becomes

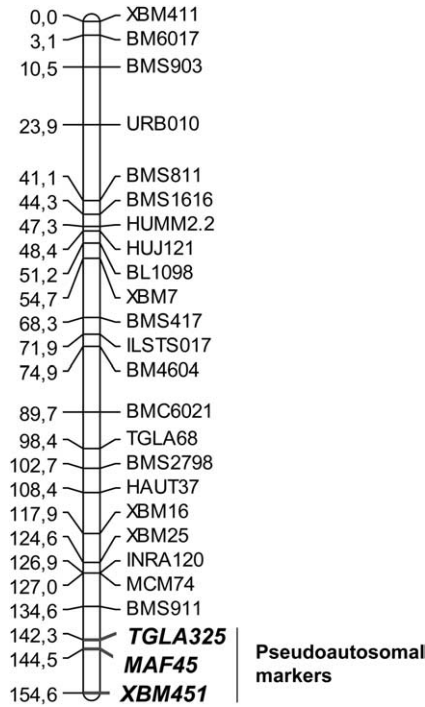


FIGURE 1.—Order and distance in centimorgans (Kosambi) between the BTAX microsatellite markers used in this study. Pseudoautosomal markers are indicated.

$$\mathbf{M} = \begin{pmatrix} \frac{1}{4}(1 - \frac{1}{2N_m})(1 - \theta_m)^2 & \frac{1}{2}(1 - \theta_m)(1 - \theta_f) \\ \frac{1}{4}(1 - \frac{1}{2N_m})(1 - \theta_m)^2 & \frac{1}{2}(1 - \theta_m)(1 - \theta_f) \\ \frac{1}{4}(1 - \frac{1}{2N_m})(1 - \theta_m)^2 & \frac{1}{2}(1 - \theta_m)(1 - \theta_f) \\ 0 & 0 \\ \frac{1}{4}(1 - \frac{1}{2N_f})(1 - \theta_f)^2 & \frac{1}{4}(\frac{1}{2N_m}(1 - \theta_m)^2 + \frac{1}{2N_f}(1 - \theta_f)^2) \\ \frac{1}{4}(1 - \frac{1}{2N_f})(1 - \theta_f)^2 & \frac{1}{4}(\frac{1}{2N_m}(1 - \theta_m)^2 + \frac{1}{2N_f}(1 - \theta_f)^2) \\ \frac{1}{4}(1 - \frac{1}{2N_f})(1 - \theta_f)^2 & \frac{1}{4}(\frac{1}{2N_m}(1 - \theta_m)^2 + \frac{1}{2N_f}(1 - \theta_f)^2) \\ 0 & 1 \end{pmatrix} \quad (4)$$

The probabilities of coalescence without recombination, p_C^T , computed as described above were then utilized to compute IBD probabilities for pairs of X chromosomes conditional on flanking marker genotypes, Φ_p , using the rules defined by MEUWISSEN and GODDARD (2001).

Mapping QTL on the X chromosome using a restricted maximum-likelihood approach: QTL were mapped on the X chromosome using an “interval-mapping approach”; *i.e.*, the position of the hypothetical QTL was slid across the chromosome X (BTAX) marker map (Figure 1) and lod scores were computed at regular intervals. In this study, a lod score was generated in the middle of each marker interval.

Lod scores were computed essentially as previously described (KIM and GEORGES 2002; BLOTT *et al.* 2003). To test map position p , we first clustered the n BTAX haplotypes in the data set in a rooted dendrogram. The $n(n-1)/2$ pairwise $(1 - \Phi_p)$ values (computed as described above) were used as distance measures and UPGMA as a hierarchical clustering algorithm (*e.g.*, MOUNT 2001). The tree was then used as a logical framework to group the haplotypes in functionally distinct clusters. To achieve this, the tree was scanned downward from the top and

branches cut such that the distance between haplotypes within resulting subtrees would not exceed a threshold C .

We then modeled the breeding values of the sons using the following linear model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_h\mathbf{h} + \mathbf{Z}_u\mathbf{u} + \mathbf{e}. \quad (5)$$

$\mathbf{y}$ ($n \times 1$) is the vector of breeding values for all sons. $\mathbf{b}$ ($p \times 1$) is a vector of fixed effects, which reduces here to the overall mean. $\mathbf{X}$ ($n \times p$) is the incidence matrix relating fixed effects to individual sons, which here reduces to a vector of ones. $\mathbf{h}$ ($q \times 1$) is the vector of random QTL effects corresponding to the functionally distinct haplotype clusters defined as above. $\mathbf{Z}_h$ ($n \times q$) is an incidence matrix that relates each son to his corresponding haplotype cluster. $\mathbf{u}$ ($r \times 1$) is the vector of random individual autosomal polygenic effects (“animal model,” LYNCH and WALSH 1998). $\mathbf{Z}_u$ ($n \times r$) is a diagonal incidence matrix that relates each son to its individual polygenic effect. $\mathbf{e}$ ($n \times 1$) is the vector of individual error terms.

Haplotype cluster effects with corresponding variance, σ_H^2 , individual polygenic effects with corresponding variance, σ_A^2 , and individual error terms with corresponding variance, σ_E^2 , were estimated using AIREML (JOHNSON and THOMPSON 1995), by maximizing the restricted log-likelihood function L :

$$L = -0.5 \ln |\mathbf{V}| - 0.5 \ln |\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}| - 0.5 (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}). \quad (6)$$

In this,

$$\mathbf{V} = \sigma_H^2 \mathbf{Z}_h \mathbf{H} \mathbf{Z}_h^T + \sigma_A^2 \mathbf{Z}_u \mathbf{A} \mathbf{Z}_u^T + \sigma_E^2 \mathbf{I}. \quad (7)$$

Because we assumed that the covariance between the QTL effects of the different haplotype clusters was zero, $\mathbf{H}$ reduces to an identity matrix. $\mathbf{A}$ is the additive genetic relationship matrix (LYNCH and WALSH 1998).

L was computed for all possible values of C (from 1 to 0), to identify a restricted maximum-likelihood (REML) solution for map position p under the H_1 hypothesis.

The log-likelihood of the data under H_1 was then compared with that under the null hypothesis, H_0 , of no QTL at map position p . The latter was computed as described above but using the reduced model:

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_u\mathbf{u} + \mathbf{e}. \quad (8)$$

Evidence in favor of a QTL at map position p was expressed as a lod score:

$$z_p = \frac{(L_{H_1} - L_{H_0})}{\ln(10)}. \quad (9)$$

The statistical significance of observed z_p -values, accounting for the fact that we tested 24 marker intervals and multiple values of C in each interval, was estimated from the distribution of largest z_p -values obtained along the X chromosome when analyzing 1000 sets of simulated phenotypes. The latter were generated by “dropping” 30 unlinked, autosomal QTL through the known genealogy of the 929 bulls. All QTL were assumed to have four alleles with random allelic frequency and additive QTL effects sampled from $e^{-x}/2$ ($x \in [0; +\infty]$) with randomly assigned positive or negative sign. The polygenic effect of each bull was computed as the sum of its 60 QTL effects. In addition, each bull was assigned a nongenetic effect sampled from a normal distribution $N(0, \sigma_E^2)$ such that

$$\sigma_E^2 = \frac{\sigma_F^2(1 - h^2)}{h^2}, \quad (10)$$

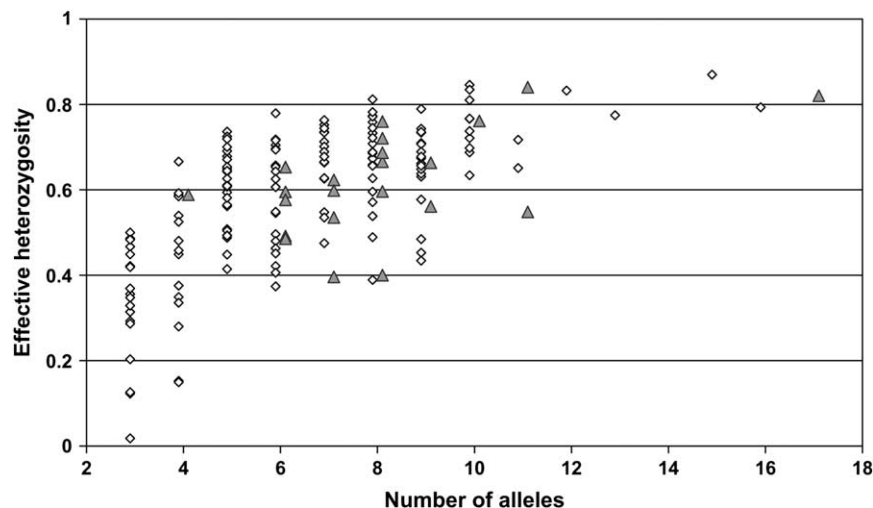


FIGURE 2.—Number of alleles and effective heterozygosity ($1 - \sum_{i=1}^n p_i^2$) measured for 202 autosomal (open diamonds) and 25 X-specific (shaded triangles) microsatellite markers in a Holstein–Friesian granddaughter design.

where σ_p^2 is the variance due to the polygenic effects and h^2 is the heritability that was set at 0.75.

RESULTS

A total of 276,048 phased autosomal and 22,181 X-linked microsatellite genotypes, produced on 929 bulls as described in MATERIALS AND METHODS, were available for analysis. Figure 2 compares the number of alleles as well as the “effective” heterozygosity (computed as $1 - \sum_{i=1}^n p_i^2$, where p_i is the frequency of allele i of n) for the 202 autosomal microsatellites and the 25 gonosomal microsatellites. Only the alleles inherited from the dam were considered in this analysis. The average number of alleles and effective heterozygosity ($\pm$ SE) were, respectively, 6.6 ± 0.4 and 0.59 ± 0.02 for the autosomal markers *vs.*, respectively, 8.0 ± 1.1 and 0.61 ± 0.05 for the gonosomal markers. Microsatellite markers appeared thus slightly but not significantly more polymorphic on the X chromosome than on the autosomes.

Figure 3A compares the level of pairwise LD measured in the same sample for syntenic autosomal and X-specific markers as a function of genetic distance. To allow for proper comparison, the distance between markers was set at $\theta_A = \frac{1}{2}(\theta_m + \theta_f)$ for the autosomal markers and at $\theta_X = [\theta_f(N_m + N_f)]/[N_m + 2N_f]$ for the gonosomal markers. In these, θ_m and θ_f correspond to the male and female recombination rates, while N_m and N_f correspond to the numbers of males and females, respectively. For the autosomes, θ_A corresponded to the sex-averaged θ -values obtained from IHARA *et al.* (2004). For the X chromosome, θ_f -values were obtained from SONSTEGARD *et al.* (2001). N_m and N_f were set at 25 and 1 million, respectively, yielding an effective (autosomal) population size (N_{eA}) of ≈ 100 . The latter value is close to recent, pedigree-based estimates of N_{eA} in this population (BOICHARD 1996). It can be seen that r^2 -values are systematically higher for the X chromosome than

for the autosomes, being ~ 1.75 times as large and ranging from values of the order of 0.10 for recombination rates $< 5\%$ to values of the order of 0.01 for recombination rates between 20 and 25%. Note that these r^2 -values, including those between the more distant markers, are highly significant (FARNIR *et al.* 2000). Note also that choosing less extreme sex ratios accentuates the difference between r_X^2 and r_A^2 .

Using the Markov model described in MATERIALS AND METHODS, we computed the $(929^2 - 929)/2$ pairwise IBD probabilities conditional on marker data, for the 24 BTAX marker intervals. In these calculations, we assumed that T , the number of generations to the base population, was 100 and that N_m was 25 and N_f 1 million—as before. The overall IBD probability (averaged across all chromosome pairs and marker intervals) was 0.26. This is lower than the theoretical expectation of 0.36, corresponding to the probability of coalescence of a chromosome region of size 0 computed using the gonosomal-specific Markov model (see MATERIALS AND METHODS), and hence to the coefficient of kinship of a chromosome region of size 0 (MEUWISSEN and GODDARD 2001). Adjusting these parameter values for the theoretical and observed values to match might be a way to fine tune the model.

For a randomly selected interval (interval 12), we compared the IBD probabilities obtained with the X-specific transition probabilities, with those that would be obtained using autosomal transition probabilities, assuming (i) 25 males and 1 million females, but sex-specific recombination rates (θ_f and $\theta_m = 0$), and (ii) a population size of 100 with equal numbers of males and females, as well as equal recombination rates in males and females ($= \frac{1}{2}\theta_f$). The latter yields in essence what would be obtained when directly applying the autosomal method proposed by MEUWISSEN and GODDARD (2001) to the X chromosome. The results obtained under scenario i are shown in Figure 4. It can be seen that—although the differences are quite modest—the

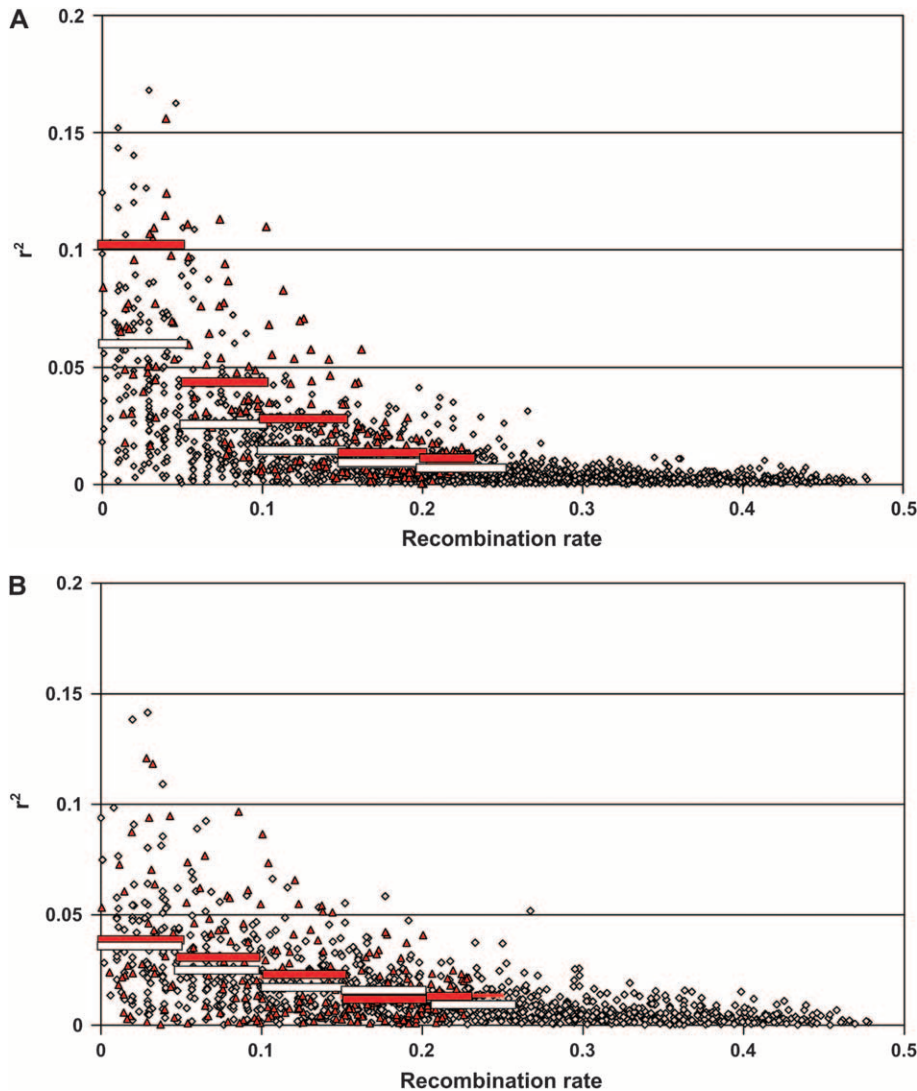


FIGURE 3.—Linkage disequilibrium [(A) real data and (B) simulated data] measured using r^2 for syntenic autosomal (white diamonds) and X-specific (red triangles) microsatellite markers. White and red horizontal bars correspond to average autosomal and X-specific r^2 -values for windows of 5 recombination units (0–5%, 5–10%, 10–15%, ...).

correct gonosomal model yields more “conservative” IBD probabilities, in the sense that if marker information suggests that the considered pair of chromosomes has a higher probability of IBD than two randomly

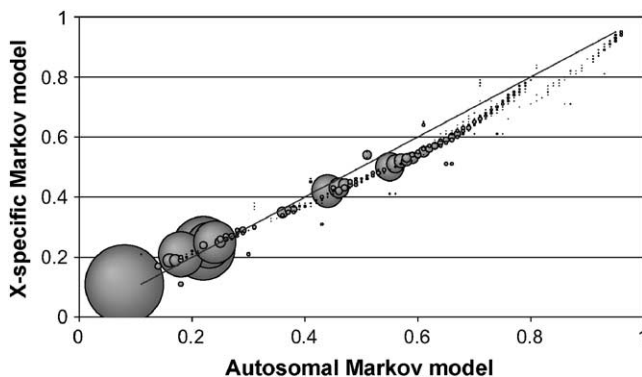


FIGURE 4.—Comparison of IBD probabilities computed using either the X-specific (y-axis) or the autosomal Markov model (x-axis). The size of the symbol reflects the number of corresponding observations. The line corresponds to the expectations if both models were yielding the same probabilities.

selected chromosomes this probability is inflated when using the erroneous autosomal model, while if the marker information suggests the opposite, the autosomal model underscores the IBD probabilities when compared to the correct gonosomal model. This tendency was even more pronounced when using an autosomal model assuming equal numbers of males and females and equal recombination rate in both sexes corresponding to $\frac{1}{2}\theta_f$ (scenario ii; data not shown).

The pairwise IBD probabilities described in the previous section were utilized to search for QTL influencing 48 dairying traits on the X chromosome using the variance component approach described in MATERIALS AND METHODS. The proposed model includes a random individual animal effect that should properly correct for differences in autosomal coancestry that may be correlated with BTAX IBD probabilities and thus protect against false-positive QTL effects as a result of stratification. The lod score threshold of 2.5 corresponding to a chromosomewide type I error of $\approx 5\%$ was exceeded for five traits: fat yield, direct durability, durable prestation, milking speed, and rear leg set (Table 2). Assuming that

TABLE 2
QTL mapped on BTAX

Trait	Marker bracket	Chromosomewide <i>P</i> -value	No. clusters	r_H^2	r_A^2	r_E^2
Fat yield	[HAUT37-XBM16]	0.038	487	0.08	0.80	0.12
		0.042	499	0.08	0.80	0.12
Direct durability	[ILSTS017-BM4604]	0.050	745	0.08	0.90	0.02
		0.051	741	0.08	0.90	0.02
Durable prestation	[TGLA325-MAF45]	0.051	2	0.07	0.86	0.07
		0.053	2	0.06	0.85	0.08
Milking speed	[XBM7-BMS417]	0.051	64	0.03	0.95	0.02
		0.053	76	0.02	0.96	0.03
Rear leg set	[BMS1616-HUMM2.21]	0.050	23	0.03	0.96	0.01
		0.052	14	0.04	0.96	0

The first line in each cell shows values obtained when computing IBD probabilities using $N_m = 25$, $N_f = 1,000,000$, and $T = 100$, and the second line shows the corresponding values obtained using $N_m = 15,000$, $N_f = 1,000,000$, and $T = 100$. σ_A^2 and σ_E^2 correspond, respectively, to the estimated proportion of the trait variance explained by the BTAX QTL, the autosomal polygenes and the error term. No. of clusters shows the number of haplotype clusters yielding the model that maximizes the likelihood (REML) of the data. The high values of r_A^2 are explained by the fact that the analyzed traits are estimated breeding values.

these are independent QTL (as suggested by the type of trait and the respective most likely QTL positions), and knowing that—because of their correlations—the 48 analyzed traits behave in essence as ≈ 20 independent traits (W. COPPIETERS, unpublished data), this is thus approximately five times more than expected by chance alone. The same five QTL emerged out of an analysis performed using less extreme parameter values $N_m = 15,000$, $N_f = 1,000,000$, and $T = 100$ (Table 1). Altogether, this suggests that at least some of these QTL are likely to be genuine.

The proportion of the trait variance explained by the QTL ranged from 2 to 8% and was—as expected—well correlated with the lod score value. The number of haplotype clusters associated with the REML solution ranged very widely from 2 to 745. Figure 5 shows representative examples of the distribution of haplotype-cluster effects with standard error of the estimates and corresponding cluster frequency in the analyzed sample. In all cases, we were able to identify common haplotype clusters with significantly different effects. These most likely explain the high lod scores that were obtained.

DISCUSSION

In this article, we have compared the level of polymorphism and LD between X-linked and autosomal microsatellite markers in the Holstein–Friesian dairy cattle population. We find X-linked microsatellite markers to be at least as variable as autosomal ones and—at comparable distance—LD to be considerably higher on the X chromosome than on the autosomes.

Typically, genetic polymorphism is expected to be lower and LD to be higher for markers on the X chromosome, as observed in humans (*e.g.*, DIB *et al.* 1996; SACHIDANANDAM *et al.* 2001; SCHAFFNER 2004; ALTSHULER *et al.* 2005). The lower level of polymorphism on the X is thought to result from (i) higher genetic drift as $N_{eX} = \frac{3}{4}N_{eA}$, (ii) a reduced mutation rate as the male mutation rate (μ_M) is expected to be 1.7 to 4 times higher than the female mutation rate (μ_F), and $\mu_X = \frac{2}{3}\mu_F + \frac{1}{3}\mu_M$ while $\mu_A = \frac{1}{2}\mu_F + \frac{1}{2}\mu_M$ (*e.g.*, ELLEGREN 2000; SACHIDANANDAM *et al.* 2001), and (iii) enhanced purifying selection due to male hemizyosity. Note that ascertainment bias (*i.e.*, the fact that one selects on polymorphism when developing markers) is expected to considerably blur this picture. The higher level of LD on the X chromosome is primarily thought to result from higher genetic drift. Our results in the bovine may thus—at first glance—seem in reasonable agreement with theoretical predictions: the level of polymorphism on the X is comparable to that on autosomes as a result of ascertainment bias when developing microsatellite markers, but the higher level of LD on the X—which is not subject to the same ascertainment bias—is as predicted by basic population genetics.

As a matter of fact, $N_{eX} = \frac{3}{4}N_{eA}$ is valid only when the sex ratio ($SR = N_m/N_f$) is one. This was clearly not the case for domestic cattle and most probably was not for their wild-type ancestors. Assuming (*e.g.*, WRIGHT 1969) that

$$N_{eA} = \frac{4N_mN_f}{N_m + N_f} \quad (11)$$

and

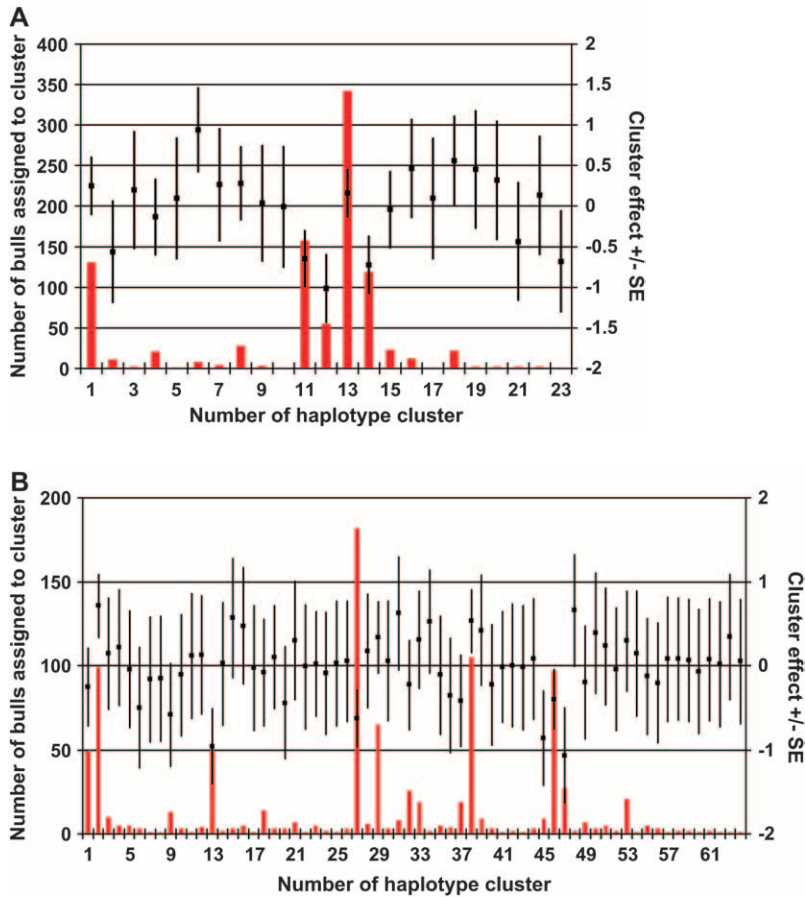


FIGURE 5.—Representative examples of population frequency (red bars) and effect $\pm$ standard error for the haplotype clusters yielding the highest lod scores for rear leg set (A) and milking speed (B).

$$N_{ex} = \frac{9N_m N_f}{4N_m + 2N_f} \quad (12)$$

one can easily show that N_{ex} becomes larger than N_{eA} when $SR < 1/7$ (Figure 6). Thus in cattle N_{ex} is likely to have been higher than N_{eA} before and certainly after domestication. Equations 11 and 12 assume that all

parents of a given sex have an equal chance of contributing offspring to the next generation, which is obviously not valid in livestock. Accounting for non-random parental contribution, however, does not alter these conclusions (Figure 6 and APPENDIX).

This may in part explain why—in the bovine—X-linked markers are as polymorphic as autosomal markers,

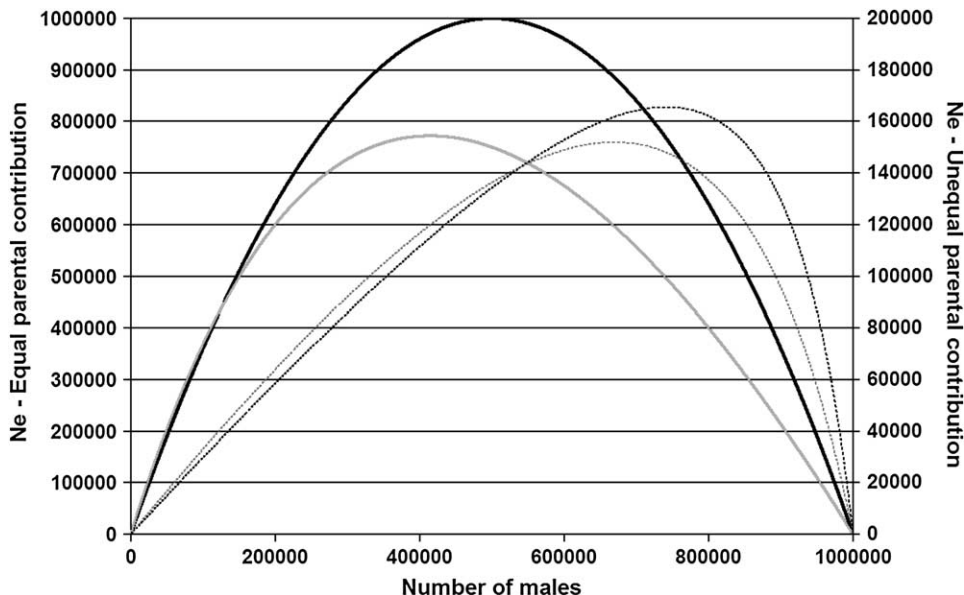


FIGURE 6.—Estimates of autosomal (black) and X-specific (gray) effective population size for a population of constant size (1 million individuals) but with varying sex ratio assuming (i) equal probability of parenthood for all parents (solid lines, left y-axis) or (ii) unequal probability of parenthood among parents (stippled lines, right y-axis). For the latter scenario we assumed that the males were subdivided in three categories representing, respectively, 0.001, 0.1, and 0.89 of the male population with relative probabilities of parenthood of 100, 10, and 1, while the females were subdivided in three categories representing, respectively, 0.001, 0.1, and 0.89 of the female population with relative probabilities of parenthood of 10, 2, and 1.

contrary to what is observed in humans. However, it precludes stronger drift from underlying the higher level of LD between X-linked markers. This is at least in part corroborated by comparing the level of LD between autosomal and X-linked “*in silico*” markers whose segregation within the known genealogy of the 929 analyzed bulls is simulated by gene dropping as previously described (FARNIR *et al.* 2000). While for autosomal markers *in silico* LD levels are of the same magnitude as actual LD levels, for X-linked markers the levels of *in silico* LD are of the same magnitude as for the autosomal markers and thus considerably lower than the actual LD levels (Figure 3B).

This strongly suggests that the higher level of LD observed on the X chromosome of cattle, and possibly other species as well, is due to as yet undetermined factors other than drift. The fact that the increased levels of LD are also observed for markers that are >10 cM apart suggests that these factors are still operating or have been operating until recently, certainly after domestication (*e.g.*, HAYES *et al.* 2003).

We have developed a novel approach based on a Markovian model that allows computation of IBD probabilities conditional on flanking marker data accommodating both autosomes and gonosomes, varying SR, and sex-specific recombination rates. At present, the model accounts only for recombination but it should be possible to extend it such that it includes mutation as well. This may be important when dealing with short chromosomal segments for which mutation and recombination rate are of the same order of magnitude.

It is worthwhile noting that it should be possible to use the variance of IBD probabilities (σ_{OBS}^2) as the basis for a measure of LD information content (IC). For a given marker interval, σ_{OBS}^2 could be compared with the variance expected if the marker information allowed one to unambiguously distinguish IBD status from non-IBD status. The variance corresponding to such an ideal map is

$$\sigma_{\text{MAX}}^2 = \bar{x}(1 - \bar{x})^2 + (1 - \bar{x})\bar{x}^2 = \bar{x}(1 - \bar{x}), \quad (13)$$

where $\bar{x}$ corresponds to the average IBD probability in the considered interval. Indeed, a proportion $\bar{x}$ of chromosome pairs are expected to be IBD, thus have an IBD probability of 1 if the markers are fully informative, and hence contribute $(1 - \bar{x})^2$ to the variance of IBD probabilities; a proportion $(1 - \bar{x})$ of chromosome pairs are expected to be non-IBD, thus have an IBD probability of 0 if the markers are fully informative, and hence contribute $\bar{x}^2$ to the variance of IBD probabilities. The ratio $\sigma_{\text{OBS}}^2/\sigma_{\text{MAX}}^2$ could thus be viewed as a measure of the LD IC of the utilized map. Although the detailed behavior and suitability of the proposed IC measure still need to be examined in more detail, the IC profile obtained with this method along the X chromosome map (data not shown) suggests—as expected—that one could gain considerable power by increasing the marker density. The recent availability of tens of thousands of

bovine single-nucleotide polymorphisms (SNPs) as well as array-based methods allowing for the effective genotyping now makes this possible for both autosomes and the X chromosome.

We present evidence in this work that the X chromosome harbors several QTL affecting traits of importance to the dairy breeding industry. This contradicts previous studies in which effects of the X chromosome on some of these traits were searched for by fitting models in which the expected phenotypic covariance between relatives included terms accounting for sex linkage (*e.g.*, LYNCH and WALSH 1998; M. GODDARD, personal communication). The discrepancy between these results might be due to substantial gains of detection power from considering marker information.

Contrary to MEUWISSEN and GODDARD (2001), but following KIM and GEORGES (2002) and BLOTT *et al.* (2003), we are not directly converting between haplotype IBD probabilities in covariances between haplotype effects, but are clustering haplotypes on the basis of the IBD probabilities and then testing the effect of haplotype clusters on phenotype. One of the advantages of this approach is that it allows one to identify the haplotypes that are the most likely to be functionally distinct (as illustrated in Figure 5) and then to focus the molecular work on these haplotypes when searching for causal variants. In addition, the haplotype clustering method in essence models a discrete number of QTL alleles that may be closer to the actual biology than the infinite number of alleles that is modeled in the alternative approach.

Our results suggest that it might be advantageous to include X-linked markers when performing marker-assisted selection or genomic selection (MEUWISSEN *et al.* 2001) in dairy cattle. One could, for instance, select among full brothers according to the X chromosome inherited from the dam.

We are grateful for the expert technical assistance of Paulette Berzi, Nadine Cambisano, Latifa Karim, Myriam Mni, Patricia Simon, and Erica Davis in producing the microsatellite genotypes and to Alain Empain for expert administration of the computer grid. This work has been funded by grants from Holland Genetics (Arnhem, The Netherlands), the Livestock Improvement Corporation (Hamilton, New Zealand), the Walloon Ministry of Agriculture, and the Genetic Analysis of Multifactorial Entities ULg/Action de Recherche Concertée. Cynthia Sandor is a fellow of the Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture.

LITERATURE CITED

- ALTSHULER, D., L. D. BROOKS, A. CHAKRAVARTI, F. S. COLLINS, M. J. DALY *et al.*, 2005 A haplotype map of the human genome. *Nature* **437**: 1299–1320.
- BLOTT, S., J. J. KIM, S. MOISIO, A. SCHMIDT-KUNTZEL, A. CORNET *et al.*, 2003 Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* **163**: 253–266.
- BOICHARD, D., 1996 Analyse généalogique des races bovines laitières françaises. *INRA Prod. Anim.* **9**: 323–335.

- CABALLERO, A., 1994 Developments in the prediction of effective population size. *Heredity* **73**: 657–679.
- COPPIETERS, W., J. RIQUET, J.-J. ARRANZ, P. BERZI, N. CAMBISANO *et al.*, 1998 A QTL with major effect on milk yield and composition maps to bovine chromosome 14. *Mamm. Genome* **9**: 540–544.
- DIB, C., S. FAURÉ, C. FIZAMES, D. SAMSON, N. DROUOT *et al.*, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**: 152–154.
- ELLEGREN, H., 2000 Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat. Genet.* **24**: 400–402.
- FARNIR, F., W. COPPIETERS, J.-J. ARRANZ, P. BERZI, N. CAMBISANO *et al.*, 2000 Extensive genome-wide linkage disequilibrium in cattle. *Genome Res.* **10**: 220–227.
- FARNIR, F., B. GRISART, W. COPPIETERS, J. RIQUET, P. BERZI *et al.*, 2002 Simultaneous mining of linkage and linkage disequilibrium to fine-map QTL in outbred half-sib pedigrees: revisiting the location of a QTL with major effect on milk production on bovine chromosome 14. *Genetics* **161**: 275–287.
- GEORGES, M., D. NIELSEN, M. MACKINNON, A. MISHRA, R. OKIMOTO *et al.*, 1995 Mapping quantitative trait loci controlling milk production by exploiting progeny testing. *Genetics* **139**: 907–920.
- GRISART, B., F. FARNIR, L. KARIM, N. CAMBISANO, J. J. KIM *et al.*, 2004 Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc. Natl. Acad. Sci. USA* **101**: 2398–2403.
- HAYES, B. J., P. M. VISSCHER, H. C. MCPARTLAN and M. E. GODDARD, 2003 Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* **13**: 635–643.
- HIENDLEDER, S., H. THOMSEN, N. REINSCH, J. BENNEWITZ, B. LEYHE-HORN *et al.*, 2003 Mapping of QTL for body conformation and behavior in cattle. *J. Hered.* **94**: 496–506.
- IHARA, N., A. TAKASUGA, K. MIZOSHITA, H. TAKEDA, M. SUGIMOTO *et al.*, 2004 A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Res.* **14**: 1987–1998.
- JOHNSON, D. L., and R. THOMPSON, 1995 Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.* **78**: 448–456.
- KIM, J. J., and M. GEORGES, 2002 Evaluation of a new fine-mapping method exploiting linkage disequilibrium: a case study analysing a QTL with major effect on milk composition on bovine chromosome 14. *Asian-Aust. J. Anim. Sci.* **15**: 1250–1256.
- KUHN, C. H., J. BENNEWITZ, N. REINSCH, N. XU, H. THOMSEN *et al.*, 2003 Quantitative trait loci mapping of functional traits in the German Holstein cattle population. *J. Dairy Sci.* **86**: 360–368.
- LYNCH, M., and B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.
- MEUWISSEN, T. H., and M. E. GODDARD, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**: 421–430.
- MEUWISSEN, T. H., and M. E. GODDARD, 2001 Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.* **33**: 605–634.
- MEUWISSEN, T. H., B. J. HAYES and M. E. GODDARD, 2001 Prediction of total genetic value using genomewide dense marker maps. *Genetics* **157**: 1819–1829.
- MOUNT, D. W., 2001 *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- SACHIDANANDAM, R., D. WEISSMAN, S. C. SCHMIDT, J. M. KAKOL, L. D. STEIN *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
- SCHAFFNER, S. F., 2004 The X chromosome in population genetics. *Nat. Rev. Genet.* **5**: 43–51.
- SONSTEGARD, T. S., W. BARENDSE, G. L. BENNETT, G. A. BROCKMANN, S. DAVIS *et al.*, 2001 Consensus and comprehensive linkage maps of the bovine sex chromosomes. *Anim. Genet.* **32**: 115–117.
- WRIGHT, S., 1969 The theory of gene frequencies, p. 213 in *Evolution and the Genetics of Populations*, Vol. 2. University of Chicago Press, Chicago/London.

Communicating editor: C. HALEY

APPENDIX

In the case of nonrandom parental contribution and unequal numbers of males and females, effective population size for autosomes (N_{e_A}) and the X chromosome (N_{e_X}) can be estimated using, respectively,

$$\frac{1}{N_{e_A}} = \frac{1}{16N_m} [2 + \sigma_{mm}^2 + 2SR\rho_{mm,mf} + SR^2\sigma_{mf}^2] + \frac{1}{16N_f} \left[2 + \sigma_{ff}^2 + \frac{2}{SR}\rho_{fm,ff} + \frac{1}{SR^2}\sigma_{fm}^2 \right] \quad (A1)$$

and

$$\frac{1}{N_{e_X}} = \frac{1}{9N_m} [1 + 2SR^2\sigma_{mf}^2] + \frac{1}{9N_f} \left[1 + \sigma_{ff}^2 + \frac{2}{SR}\rho_{fm,ff} + \frac{1}{SR^2}\sigma_{fm}^2 \right]. \quad (A2)$$

In these equations N_m , N_f , and SR are as defined before, σ_{xy}^2 is the variance of the number of offspring of sex y per parent of sex x , and $\rho_{xy,xz}$ is the covariance between the number of offspring of sex y and the number of offspring of sex z per parent of sex x (*e.g.*, CABALLERO 1994).

Unequal parental contribution can be modeled by assuming that the N_m males belong to K classes of size N_{m_i} with probability of fatherhood p_{m_i} , such that $\sum_{i=1}^K N_{m_i}p_{m_i} = 1$, and likewise that the N_f females belong to L classes of size N_{f_i} with probability of motherhood p_{f_i} , such that $\sum_{i=1}^L N_{f_i}p_{f_i} = 1$. Using this model, one can show that

$$\begin{aligned} \sigma_{mm}^2 &\approx N_m \sum_{i=1}^K N_{m_i} p_{m_i}^2 \\ \sigma_{mf}^2 &\approx \frac{1}{SR^2} (\sigma_{mm}^2 - 1) + \frac{1}{SR} \\ \sigma_{ff}^2 &\approx N_f \sum_{i=1}^L N_{f_i} p_{f_i}^2 \\ \sigma_{fm}^2 &\approx SR^2 (\sigma_{ff}^2 - 1) + SR \\ \rho_{mm,mf} &\approx \frac{1}{SR} (\sigma_{mm}^2 - 1) \\ \rho_{fm,ff} &\approx SR (\sigma_{ff}^2 - 1). \end{aligned}$$

We supposed that the effects of selection were independent from one generation to another; offspring do not inherit probability of parenthood.

Figure 6 compares estimates of N_{e_A} and N_{e_X} obtained with this model in a population of 1 million individuals and varying SR in the case of (i) equal probability of parenthood for all parents and (ii) unequal probability of parenthood between parents. In the latter case, it was assumed that (i) the males were subdivided in three categories representing, respectively, 0.001, 0.1, and 0.89 of the male population with relative probabilities of parenthood of 100, 10, and 1 and (ii) the females were subdivided in three categories representing, respectively, 0.001, 0.1, and 0.89 of the female population with relative probabilities of parenthood of 10, 2, and 1.

We used this model to estimate N_{e_A} and N_{e_X} in the Dutch Holstein–Friesian population. Using statistics for 2002, 2003, and 2004 obtained from the Nationaal Rundveeteelt Syndikaat (Arnhem, The Netherlands), we set the number of reproducing males at 15,650 and the number of reproducing females at 1,076,100. To match the known distributions of sons per bull–sire and bull–dam as well as daughters per son as closely as possible (data not shown), we subdivided (i) the reproducing males in five classes representing, respectively, 0.0005, 0.005, 0.2, 0.3, and 0.4945 of the male population with relative probabilities of parenthood of 15,000, 400, 10, 2, and 1 and (ii) the reproducing females in three classes representing, respectively, 0.05, 0.1, and 0.85 of the female population with relative probabilities of parenthood of 1000, 100, and 1. This yielded estimates of N_{e_A} and N_{e_X} of 88 and 99, respectively, thus very similar to the corresponding values of 100 and 112 obtained when considering 25 reproducing males and 1,000,000 reproducing females with equal probability of parenthood (as utilized throughout this article), as well as to pedigree-based estimates of N_{e_A} (BOICHARD 1996).