Quantitative Trait Locus Mapping in Dairy Cattle by Means of Selective Milk DNA Pooling Using Dinucleotide Microsatellite Markers: Analysis of Milk Protein Percentage

Ehud Lipkin,* Mathias O. Mosig,* Ariel Darvasi,*,[1] Ephraim Ezra,† A. Shalom,* Adam Friedmann* and Morris Soller*

**Department of Genetics, Alexander Silberman Institute of Life Science, Hebrew University of Jerusalem, Jerusalem 91904, Israel and †Israel Cattle Breeder's Association, Caesaria, Israel*

Manuscript received March 25, 1997
Accepted for publication April 13, 1998

ABSTRACT

"Selective DNA pooling" accomplishes quantitative trait locus (QTL) mapping through densitometric estimates of marker allele frequencies in pooled DNA samples of phenotypically extreme individuals. With poly(TG) microsatellites, such estimates are confounded by "shadow" ("stutter") bands. A correction procedure was developed on the basis of an observed linear regression between shadow band intensity and allele TG repeat number. Using this procedure, a selective DNA pooling study with respect to milk protein percentage was implemented in Israel-Holstein dairy cattle. Pools were prepared from milk samples of high and low daughters of each of seven sires and genotyped with respect to 11 markers. Highly significant associations with milk protein percentage were found for 5 of the markers; 4 of these markers confirmed previous reports. Selective DNA pooling accessed 80.6 and 48.3%, respectively, of the information that would have been available through individual selective genotyping or total population genotyping. In effect, the statistical power of 45,600 individual genotypings was obtained from 328 pool genotypings. This methodology can make genome-wide mapping of QTL accessible to moderately sized breeding organizations.

THE high costs of screening large populations for marker allele frequencies can be reduced by the use of "selective DNA pooling." In this procedure, determination of linkage between a molecular marker and a quantitative trait locus (QTL) is based on the distribution of parental marker alleles among pooled DNA samples of the extreme high and low phenotypic groups of offspring (Darvasi and Soller 1994). At present, most of the markers used in constructing genomic maps are dinucleotide microsatellites having a poly(TG) motif (Barendse *et al.* 1994; Bishop *et al.* 1994; Cheng *et al.* 1995; Crawford *et al.* 1995; Dib *et al.* 1996; Dietrich *et al.* 1996). These are typed by PCR (Tautz 1989; Weber and May 1989). Alleles, differing in repeat number, are resolved as discrete bands by electrophoresis through an acrylamid sequencing gel. Estimating allele frequencies in pooled DNA samples is based on a linear relation between the initial number of copies of the allele in the pool, as determined by the allele frequency in the group of individuals making up the pool and the final allele band intensity, as determined by densitometry. However, this estimate is often confounded by the presence of "shadow" or "stutter" bands, *i.e.*, artifactual PCR products derived from the genomic poly(TG) tract by deletion or insertion of one or more repeat motifs (Hauge and Litt 1993; Litt *et al.* 1993; Murray *et al.* 1993). Consequently, an observed band in the pool can be a composite of a main product of the genomic tract and a number of shadow products generated by contiguous alleles in the same lane.

In this article we show that the relative intensity of a shadow band is a tight linear function of the repeat number of the genomic tract from which it was derived and of the number of deleted or inserted motifs by which it differs from the genomic tract. This linear function was used to correct the densitometric intensity of microsatellite bands for overlapping shadow bands (for other correction procedures see Leduc *et al.* 1995 and Perlin *et al.* 1995). The corrected intensities were used to map QTL affecting milk protein percentage, by selective DNA pooling (Darvasi and Soller 1994), using milk samples (Lipkin *et al.* 1993a).

MATERIALS AND METHODS

Microsatellite genotyping

Source and preparation of individual DNA samples: *Milk somatic cells:* Milk samples were obtained from milk-recorded Israeli-Holstein dairy cows. Somatic cell counts of the milk

Corresponding author: Ehud Lipkin, Department of Genetics, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel.
E-mail: lipkin@vms.huji.ac.il

[1] *Present address:* The Jackson Laboratory, 600 Main St., Bar Harbor, ME 04693-1500.

samples were obtained at the central milk-testing laboratory of the Israel Cattle Breeders Association (ICBA), through DNA fluorescent dyeing (Ostensson *et al.* 1988). For individual genotyping, samples were prepared as described (Lipkin *et al.* 1993a) with the following modifications: an aliquot of milk containing 10^5 somatic cells, volume ranging from 1 to 3 ml according to cell count, was washed twice by suspension in 1.5 ml saline (0.9% NaCl) and centrifugation for 10 min at 1000 × *g*. The washed samples were then subjected to one of two alternative lysing protocols: (1) The pellet was resuspended in 40 ml TE (10 mm Tris, pH 7.5, 1 mm EDTA), giving a final concentration of 250 cells/μl. Resuspended cells were lysed by incubation at 100° for 5 min, at 55° for 5 min, and again at 100° for 5 min. (2) The pellet was resuspended in 20 mm NaOH and incubated for 10 min at 100°, and the NaOH was neutralized with Tris, pH 7.5, 121 mm final concentration. Again volumes were adjusted to a final concentration of 250 cells/μl.

Hair roots: For direct genotyping, hair roots of six Israeli-Holstein cows were prepared as described (Lipkin *et al.* 1993b).

Blood leukocytes: DNA was extracted from peripheral blood leukocytes of seven Israeli-Holstein sires as described (Lipkin *et al.* 1993b).

Bovine DNA: Nine DNA samples from the international bovine family panel (Barendse *et al.* 1994) were included in this study.

Preparation of pooled milk sample: In making up the pools, raw milk aliquots containing 20,000 cells were taken from each cow. Somatic cell concentrations varied from 2000 to 20 million per ml. Hence, to ensure equal cell number from each cow in the pool, disposable tips were used and externally wiped dry after milk was taken from an individual sample and before it was added to the pool.

The pooled samples were prepared for the PCR reactions as the individual milk samples. Due to the larger volume, the saline washing was carried out in 50-ml tubes, and the cells were washed up to 10 times, until the supernatant was clear. Taking into account the number of cows included in the pool and the resultant number of cells in the final washed pellet, the final pool volume was adjusted to a concentration of 5000 cells/μl; a total of 5000 cells were taken for the reaction mixture.

PCR genotyping: Samples were denatured by heating at 94° for 4-min, and subsequently amplified in a thermocycler (MJ Research, Watertown, MA). Amplification cycles included denaturation at 94° for 60 sec; annealing for 60 sec at the temperature reported for the particular primer pair; and extension at 72° for 30 sec. After 35–40 cycles for individual or pooled milk samples, and 30 cycles for hair roots and DNA samples, an extension step was carried out for 10 min at 72°. The PCR mix contained 1.0 μl of 10× PCR buffer (750 mm Tris, pH 8.3, 200 mm KCl, and 0.1% gelatin), 2 mm $MgCl_2$, 200 μm of each dNTP, 0.5 units of Taq DNA polymerase (AB Advanced Biotechnologies, Leatherhead, UK), 0.5 μm of each primer—one of them end-labeled with $[\gamma^{35}\text{-S}]$ATP. To the PCR mixture was added 1 μl of the above-lysed cells or extracted DNA and distilled water to a total volume of 10 μl. After termination of the reaction, PCR products were separated by electrophoresis through a 5% denaturing acrylamid gel. Exposure and densitometry were carried out with a phospho-imager (FUJIX BAS1000; Fuji Photo Film Co., Tokyo).

Relative intensity of shadow bands

Statistical methods: Given the densitometric values of main and shadow bands, the observed relative intensity, RI, of a given shadow band for a given allele, was calculated as

$$\mathrm{RI}_{n,i} = D_{n,i}/D_n,$$

where *n* is the number of repeats in the native genomic tract of the allele A_n; *i* is the "order" of the shadow band, *i.e.*, the number of inserted ($i = +1$) or deleted ($i = -3$ to -1) motifs by which it was derived from the genomic tract; $\mathrm{RI}_{n,i}$ is the relative intensity of the *i*th shadow band derived from the genomic tract of A_n; D_n is the densitometric intensity of the main band derived from the genomic tract of A_n; $D_{n,i}$ is the densitometric intensity of the *i*th shadow band derived from the genomic tract of A_n.

Experiments: Densitometry measurements for the microsatellites and substrates utilized in the study (Table 1) were obtained for individual samples that were homozygous for a given allele or for which the two alleles were well separated. RI were calculated for the first three leading ($i = -3$ to -1) and the first trailing shadow bands ($i = +1$). To avoid shadow band interference, none of the shadow bands included was

TABLE 1

The shadow band study

Microsatellite[a]	Sequence of poly(TG) tract	Product length (bp)	Allele repeat number[b]	*l*[c]	Main substrate	*N*[d]
HUJ107	$(AC)_{23}$	234–236	23–24	2	DNA	18
HUJI13	$(TG)_{15}$	76–88	12–18	1	DNA	5
HUJI29	$(ATGT)_2(GT)_{13}(AT)_3$	186–204	4–13	2[e]	DNA	5
HUJII77	$(GT)_{15}(GC)_2(GT)_3$	208–230	11–16	2	Milk/DNA	97
HUJ614	$(CA)_{21}$	191–197	11–15	1[f]	DNA	13
HUJ616	$TGTA(TG)_{22}GGTG$	121–132	13–24	2	Milk/DNA	110
HUJ673	$(AC)_{10}(ATAC)_2AC$	123–125	9–10	2	Hair/DNA	33

[a] Barendse *et al.* 1994; Bishop *et al.* 1994.

[b] For complex microsatellite tracts, the repeat number is the longest consecutive dinucleotide tract.

[c] Shadow step size (difference in bp between adjacent shadow bands).

[d] Number of observations.

[e] Sometimes 1.

[f] Sometimes 2.

measured at a position of a possible overlap with the $i = -3$ to $i = +2$ shadow band of some other allele.

Two microsatellites had shadow bands separated by only 1 bp (HUJI13 and HUJ614; Table 1). Population alleles, however, were separated only by full repeats. Consequently, analysis was limited to full-repeat shadow bands, since these are the only ones that can overlap allelic main bands.

ANOVA was used to examine main effects and interactions of microsatellite and repeat number on RI values within shadow band orders $i = -1$ and $i = -2$. Only the allele repeat number, n, was found to have a significant effect. Consequently, linear regression equations on n were calculated separately for each of the above four shadow band orders, as

$$\mathrm{RI}_{n,i} = \alpha_i + b_i n,$$

where α_i and b_i are the intercept and the regression coefficient of the ith shadow band, respectively.

Allele frequency estimates in pooled DNA samples

Statistical methods: *Shadow correction:* Given the D_n values of all bands in a pool, and taking into account shadow bands, the following model was used to estimate the shadow-corrected intensity (CI) of all bands:

$$\mathrm{CI}_n = D_n - \sum_{i=-3}^{+1} (\mathrm{CI}_{n-i}\,\mathrm{RI}_{n-i,i}),$$

where n, D_n, and i are as above; CI_{n-i} is the shadow-corrected intensity of D_{n-i}; $\mathrm{RI}_{n-i,i}$ is the RI of that shadow band of A_{n-i} that overlaps the main band of A_n. This is the shadow band that is derived by deletion or insertion of i repeats from A_{n-i}. For example, if $n = 20$ and $i = -2$, then $\mathrm{RI}_{n-i,i}$ would be the RI of the shadow band derived by a deletion of two repeats from the genomic A_{22} allele. Thus, $\mathrm{CI}_{n-i}\,\mathrm{RI}_{n-i,i}$ is the estimated intensity of the ith shadow band derived from the allele A_{n-i}.

For a pool with k bands there will be k equations with k unknown CIs. These can readily be solved as a set of simultaneous equations. In preliminary experiments we found that correcting for the four shadow bands, $i = -3$, -2, -1, and $+1$, was sufficient to quantify microsatellite PCR products (Lipkin *et al.* 1996).

Total allele-products: Since the shadow-band products are derived by further amplification of the products of the genomic tract, they also represent the quantitative products of the alleles present in the pool. Consequently, the total product derived from the genomic tract consists of the product in the main band, plus the product in the derived shadow bands. Because RI increases with n, a higher proportion of the total product derived from long, as compared to short, alleles is found in the shadow bands as compared to the main band. Consequently, CI values for the main band taken alone tend to underestimate the frequency of long alleles relative to short alleles. Hence, the sum of the intensities of the shadow bands, estimated from the above RI equations, was added to the CI_n, giving

$$\mathrm{CT}_n = \mathrm{CI}_n \sum_{i=-3}^{+1} \mathrm{RI}_{n,i},$$

where CT_n is the sum of all allele products; $\mathrm{RI}_{n,0} = 1.0$, so that $\mathrm{CI}_n\mathrm{RI}_{n,0}$ is the shadow-corrected intensity of the A_n main band.

Allele frequency estimates in the pools: Given CT values for all bands in the pool, the shadow-corrected frequency estimate of allele A_n in the pool was calculated as

$$F_n = \mathrm{CT}_n / \sum_{i=-3}^{+1} \mathrm{CT}_n,$$

where $\Sigma\mathrm{CT}_n$ is the sum of CT values of all k bands in the lane.

To avoid bias due to differences in the total intensity (ΣD_{1n} and ΣD_{2n}) between two replicates of the same pool, densitometric intensities of bands in the second reaction (D'_{2n}) were adjusted according to the ratio of total densitometric products between the two reactions, as follows:

$$D'_{2n} = D_{2n}\left(\sum_{n=1}^{k} D_{1n} / \sum_{n=1}^{k} D_{2n}\right).$$

Estimates based on individual genotyping: The frequency in the pool of a microsatellite allele A_n, as estimated from N individual genotypes, was calculated as

$$F_n = O_n/2N,$$

where O_n is the number of occurrences of A_n among the genotyped animals.

Experiments: Two independent experiments in which estimates of allele frequencies in pooled DNA samples were compared to estimates obtained by individual genotyping, were carried out. Regression equations and correlations were calculated between allele frequencies as estimated from each pool and those based on the individual genotypes. Pools and individuals were compared for the overall allele frequencies, sire allele frequencies, and sire allele frequency differences between high- and low-phenotypic tails.

Constructed pools: Allele repeat numbers were determined according to the sequences of the microsatellites HUJII77 (Shalom *et al.* 1994) and HUJ616 (Shalom *et al.* 1993). Eighty milk samples of Israeli-Holstein cows were individually genotyped. Six pools of 20 cows each were then constructed for each microsatellite (12 pools in all) with known allele frequencies according to the individual genotypes. Each pool was prepared twice, each preparation was amplified twice, and means of the four amplifications were compared to the known frequencies.

Selective DNA pools: Three pairs of high and low pools, each pool consisting of milk samples from an average of 81 cows, were individually genotyped, each pair with respect to one of three different microsatellites, 420 out of 505 daughters in all.

Testing for marker-QTL linkage by selective DNA pooling

Statistical methods: *"Sire" and "dam-only" alleles:* Each pool of a half-sib family produced by a heterozygous sire contains two alleles contributed by the sire and by those dams that share alleles with him. Each of these alleles is denoted according to its repeat number: A_L has the same repeat number as the long sire and A_S the same as the short sire. All other alleles are termed "dam-only alleles."

For each sire, two pools were prepared from the daughters in each phenotypic tail: an "external pool" comprising samples from the most extreme half of the tail with respect to BV *P%* and an "internal pool" comprising samples from the remainder of the daughters. Each pool was prepared in two completely independent replicates, each replicate at a different time and usually by a different person. Thus, each sire was represented by eight pools: four high pools consisting of one external and one internal high pool, each prepared in replicate, and four low pools consisting of one external and one internal low pool, each prepared in replicate.

Significance test: A test for marker-QTL linkage for an individual sire-marker combination is most simply based on rejecting the null hypothesis, $D = 0$, where D is the difference in sire allele frequencies between the high- and low-daughter pools (Khatib *et al.* 1994). Using the normal approximation, the null hypothesis with type I error, α, is rejected if

$$Z_D = D/\mathrm{SE}(D) > Z_{1-\alpha/2},$$

where SE(D) is the standard error of D; $Z_{1-\alpha/2}$ is the ordinate of the standard normal distribution such that the area from $-\infty$ to $Z_{1-\alpha/2}$ equals $1 - \alpha/2$.

A test for marker-QTL linkage for an individual marker pooled over m sires (Weller *et al.* 1990) is based on rejecting the null hypothesis with type I error, α, if

$$\sum_{i=1}^{m} Z_{iD}^2 > \chi_\alpha^2, \quad \text{d.f.} = m.$$

Estimating D: For any given pair of external pools, the following expectations hold with respect to A_L and A_S,

$$D_{EL} = F_{EL(H)} - F_{EL(L)}, \quad D_{ES} = F_{ES(H)} - F_{ES(L)}, \quad D_{EL} = -D_{ES},$$

where D_{EL} and D_{ES} are the differences in allele frequencies between the high (H)- and low (L)-external-daughter pools for A_L and A_S, respectively; $F_{ES(H)}$ and $F_{ES(L)}$ are the estimated frequencies of A_S in the high- and low-external pools, respectively.

Where there are only two alleles in the pool, D_{EL} and D_{ES} are in complete inverse correlation with one another, and there is only one independent estimate, D_{EL} or D_{ES}, of the difference in sire allele frequency, D_E, between high- and low-external pools. However, when dam-only alleles are also segregating in the population, the correlation decreases to some extent. Consequently, the mean of these two semi-independent estimates $[D_{EL} + (-D_{ES})]/2$, carries somewhat more information than each estimate separately and was taken as the estimate of D_E.

In exactly the same manner, an estimate of the difference in sire allele frequency, D_I, between high- and low-internal pools, can be obtained.

The overall difference in sire allele frequency, D_C, between high and low daughters, based on the combined estimates from the internal and external pools, is then given by

$$D_C = (D_E + D_I)/2.$$

Estimating SE(D): Since D_L and D_S are to some extent independent, in calculating SE*(D)* for external or internal pools only and, recalling that $\mathrm{Var}(X - Y) = \mathrm{Var}X + \mathrm{Var}Y - 2\,\mathrm{Cov}\,XY$, we have

$$\mathrm{SE}^2(D_E) = [\mathrm{SE}^2(D_{EL}) + \mathrm{SE}^2(D_{ES}) - 2\,\mathrm{Cov}\,D_LD_S]/4$$

and

$$\mathrm{SE}^2(D_I) = [\mathrm{SE}^2(D_{IL}) + \mathrm{SE}^2(D_{IS}) - 2\,\mathrm{Cov}\,D_LD_S]/4,$$

and for combined pools we have

$$\mathrm{SE}^2(D_C) = [\mathrm{SE}^2(D_E) + \mathrm{SE}^2(D_I)]/4,$$

where $\mathrm{SE}^2(D_{EL}) = \mathrm{SE}^2(F_{EL(H)}) + \mathrm{SE}^2(F_{EL(L)})$ and $\mathrm{SE}^2(D_{ES}) = \mathrm{SE}^2(F_{ES(H)}) + \mathrm{SE}^2(F_{ES(H)})$ are the standard error of D_E estimates based on A_L or A_S, respectively. Similar expressions with appropriate change in subscript notation apply to internal pools. Cov D_LD_S is the covariance of the D_L and D_S estimates (see below). $\mathrm{SE}^2(F_{EL(H)}) = V_{BEL(H)} + V_{T/k}$ and $\mathrm{SE}^2(F_{EL(L)}) = V_{BEL(L)} + V_{T/k}$ are the standard error of the A_L frequency estimate in a single high- or low-external pool, respectively; similar expressions, with appropriate changes in subscript notation, apply to A_S and to internal pools, $V_{BEL(H)} = [0.25/N_{EH} + M_L(1 - M_L)/N_{EH}]/4$ and $V_{BEL(L)} = [0.25/N_{EL} + M_L(1 - M_L)/N_{EL}]/4$ are the component of the standard error derived from binomial sampling variation. $V_{BEL(H)}$ consists of two components, $0.25/N_{EH}$, derived from binomial sampling of alleles from the sire, and $M_L(1 - M_L)/N_{EH}$, derived from binomial sampling of the sire alleles from the dams. Similar expressions, with appropriate changes in subscript notation, apply to $V_{BEL(L)}$, A_S, and internal pools. The factor 4 in the denominator of the above expressions derives from the fact that sire allele frequency in a given pool is the average of the sire allele frequency in the sire and sire allele frequency in the dams.

M_L and M_S are the frequencies of A_L and A_S among the dams. When external pools only were genotyped, M_L and M_S were estimated as

$$M_L = F_{EL(H)} + F_{EL(L)} - 0.5, \quad M_S = F_{ES(H)} + F_{ES(L)} - 0.5.$$

When internal pools were also genotyped, M_L and M_S were estimated in the same manner, and pooled estimates over internal and external pools were used in all calculations. N_{EH} is the number of daughters in one external high pool; similar terms, with appropriate changes in subscript notation, would be used for low and for internal pools.

V_T is the technical error variance of estimates of sire allele frequency from pool densitometry. V_T was estimated as the standard deviation of sire allele estimates from the independent replicate preparations of each pool. Homogeneity of V_T estimates from different markers was tested by Bartlett's test (Walpole and Myers 1972). k is the number of independent replicates of each pool.

Significance testing was carried out by iteration: In the first step Cov D_LD_S was set equal to 0.0 and this value used to obtain a minimal value of SE(D_E). Using this minimal value and external pools, markers were identified that were not linked to a QTL. To account for the multiple tests, the significance level for linkage identification was set to an overall marker value of $P < 0.01$. Cov D_LD_S was then calculated on the basis of all D_{EL} and D_{ES} estimates from these markers. In the second step, this estimated value of Cov D_LD_S was used to obtain a corrected value of SE(D_E). Using this corrected value, marker-QTL linkage was again assessed, and additional markers not linked to QTL on the basis of the new, larger value of SE(D_E), were identified. A new value for Cov D_LD_S was calculated on the basis of all D_{EL} and D_{ES} estimates from all markers not linked to QTL; this new value of Cov D_LD_S was used to calculate SE(D_E) and for the significance test. This test was repeated until no additional markers dropped out of significance. The final value of Cov D_LD_S was used for the definitive calculation of SE(D) for all subsequent significance tests. Markers remaining significant after the final step were tested with internal pools, and those that remained significant on the basis of the combined pools were judged to be associated with linked QTL.

Empirical estimate of SE(D_E): As a check of the calculated estimate of SE(D_E), an empirical estimate was obtained as the standard deviation of all D_E values obtained for those markers adjudged on the above basis to be not linked to a QTL.

Experiments: The microsatellite markers used in the selective DNA pooling study (Table 2) were chosen taking into account polymorphism, distribution over the genome, and previous reports of QTL in their vicinity.

Seven elite Israeli-Holstein A.I. bulls were chosen, each the sire of more than 1800 milk recorded cows in Israeli milk-recorded herds. On the basis of breeding values for milk protein percentage (BV *P%*), a list of the highest and lowest 220 daughters was prepared for each sire. Milk samples were obtained through the active cooperation of the routine milk recording system of the ICBA during January–February 1996. At the time of collection some of the listed cows were in their dry period and others had already been culled so that eventually 2270 milk samples in all were collected. From each cow, 1.5 ml of milk was stored at −20° for individual genotyping. Pools were prepared during milk collection. In the initial marker-QTL linkage screen, the two replicates of the high- and low-external pools of each heterozygous sire were each amplified once, for a total of four PCR reactions per marker-sire combination. Significant marker-QTL associations were confirmed by genotyping all internal pools of all sires with respect to that marker. The results of the internal pools were analyzed separately, and then the results of both pools of the same tail were combined (henceforth, "combined pools") for the definitive test.

In view of the large number of tests carried out in this study, a final conclusion as to the presence of marker-QTL linkages was based on an accumulation of evidence (Lander and Kruglyak 1995), including the following: (1) an overall excess of statistically significant results at $P < 0.05$ over the number expected on chance alone; (2) concentration of significant results in some markers, while other markers show few or no significant differences; (3) concordance of internal and external pools with respect to direction and overall magnitude of differences; and (4) overall statistical significance at $P < 0.01$ for the marker as a whole, pooled over all sires.

Estimating allele substitution effect

Statistical methods (pool estimates): For estimating the magnitude of the quantitative effect associated with the marker in the population as a whole, Darvasi and Soller (1994) found it convenient to calculate genotypic group frequencies in the pool. A genotypic group is defined as all daughters that inherited the same allele (A_L or A_S) from their sire. Each half-sib pool, therefore, contained two genotypic groups. The allele substitution effect is the observed quantitative difference between the two genotypic groups, α_P. Note that the allele substitution effect is twice the genotypic group effect as defined in Darvasi and Soller (1994). As pointed out by Darvasi and Soller (1994), this difference taken over the selected tails of the population is a grossly exaggerated estimate of the actual substitution effect in the population as a whole, α_T, such that

$$\alpha_T = \alpha_P / [(i_{P/2})^2],$$

where P is the proportion selected over both tails of the population, $P/2$ at each tail, $P/4$ at each external or internal pool; $i_{P/2} = 2\,X_{P/2}/P$, the selection intensity, is the mean of an upper tail of a standard normal distribution; and $X_{P/2}$ is the ordinate of the standard normal distribution at the point $Z_{P/2}$,

$$\alpha_P = \overline{A}_L - \overline{A}_S,$$

where

$$\overline{A}_L = RF_G X_H + (1 - RF_G) X_L$$

and

$$\overline{A}_S = (1 - RF_G) X_H + RF_G X_L$$

are the respective quantitative means of all daughters receiving A_L or A_S from their sire; X_H and X_L are the BV $P\%$ means of the cows actually included in the high- and the low-external pools, respectively. These means were adjusted for recorded errors made in pool preparation, including cows sampled twice, samples added to the wrong pool, and incorrect milk volumes (and consequently an incorrect number of cells) taken from a cow in the pool. Adjustments were calculated accordingly: $RF_G = (RF_{L(H)} + RF_{S(L)})/2$ is the estimate of the relative frequency (RF) of one of the genotypic groups in one of the phenotypic tails. This was calculated as the mean of two independent frequency estimates obtained for A_L from the high and low tails, respectively, as follows:

- $RF_{L(H)} = F'_{L(H)}/(F'_{L(H)} + F'_{S(H)})$ is the estimate of the RF of all daughters in the high tail receiving A_L from their sire;
- $RF_{S(L)} = F'_{S(L)}/(F'_{L(L)} + F'_{S(L)})$ is the estimate of the RF of all daughters in the low tail receiving A_S from their sire;
- $F'_{L(H)}$, $F'_{S(H)}$, $F'_{L(L)}$, and $F'_{S(L)}$ are the estimated frequencies of all daughters receiving A_L or A_S from their sire in the high or low tail. These are obtained by subtracting the contribution of the dams to the overall frequency of A_L and A_S in the pools, calculated as

$$F'_{L(H)} = 2F_{L(H)} - M_L, \quad F'_{S(H)} = 2F_{S(H)} - M_S,$$
$$F'_{L(L)} = 2F_{L(L)} - M_L, \quad F'_{S(S)} = 2F_{S(L)} - M_S,$$

where $F_{L(H)}$ and $F_{S(H)}$ and $F_{L(L)}$ and $F_{S(L)}$ are the densitometric shadow-corrected estimates of the frequencies of A_L and A_S in the high and low pools, and M_L and M_S are as defined above.

Experiment: The allele substitution effect was calculated for all sires showing a significant test ($P < 0.01$) within the significant markers. For the three marker-sire tests confirmed by individual genotyping, substitution effects were also estimated on the basis of the individual genotypes. As is the case for pool estimates, the difference between the two genotypic groups of the sire's daughters, corrected for the selection intensity, is the substitution effect. The individual genotyping analysis did not include samples with more than two alleles (see results) and cases of false parentage identification, *i.e.*, cows that did not carry any allele of their putative sire. For two of the sires, uninformative genotypes, *i.e.*, cows having the same genotype as their sire, were also excluded. Sire two was found to carry a very rare allele of the marker INRA003; consequently, following Lagziel *et al.* (1996), all cows carrying the sire genotype were assumed to have inherited this rare allele from their sire.

RESULTS

Relative intensity of shadow bands: Figure 1 shows observations and regression line for $RI_{n,-1}$ as a function of n, pooled over all seven microsatellites, and Figure 2 shows regression lines of the RI of all four shadow bands. The regression coefficients of RI on n decreased slightly from $i = -1$ to $i = -2$, proportionately more from $i = -2$ to $i = -3$. $RI_{n,+1}$ was close to $RI_{n,-3}$, and much less than RI_{n-1}, *e.g.*, at the midpoint repeat number, $RI_{15,+1}/RI_{15,-1} = 0.34$. The total amount of shadow product was remarkable: for $n = 24$, total shadow product was twice that of the main band of the allele.

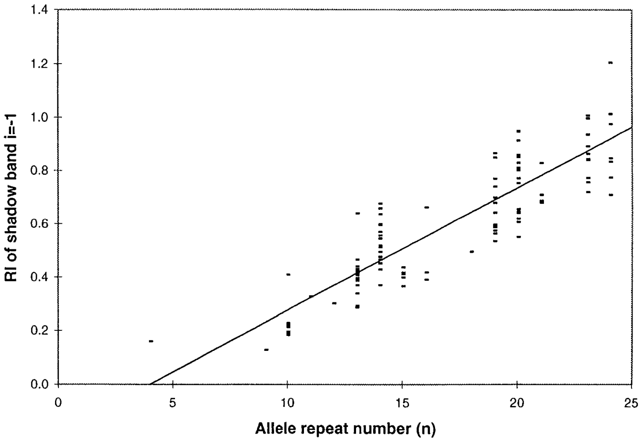

Figure 1.—$RI_{n,-1}$ of all alleles plotted against allele repeat number. ■, observation; —, regression line.

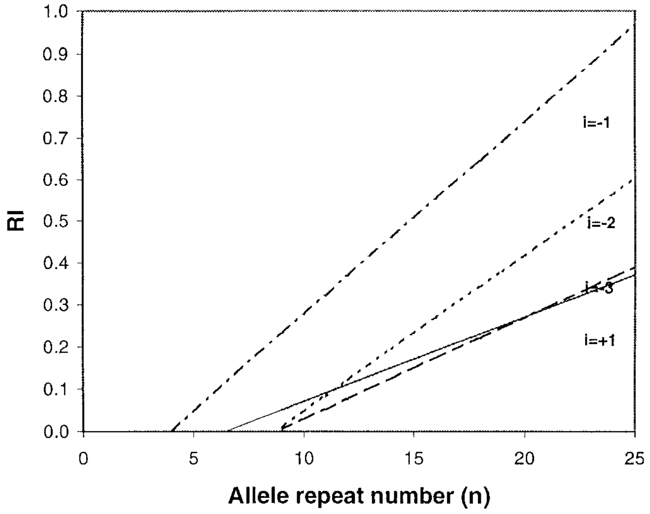

Figure 2.—Shadow band regression lines. Regression lines of RI of shadow bands on allele repeat number (n), according to shadow band order, i. N, number of observations; r, correlation coefficient; σ_e, error standard deviation; σ_a, standard error of estimate of Y-intercept; σ_b, standard error of estimate of regression coefficient.

$\mathrm{RI}_{n,-1} = -0.184 + 0.046n$, $N = 112$, $r = 0.89$, $\sigma_e = 0.108$, $\sigma_a = 0.040$, $\sigma_b = 0.002$

$\mathrm{RI}_{n,-2} = -0.323 + 0.037n$, $N = 92$, $r = 0.86$, $\sigma_e = 0.091$, $\sigma_a = 0.040$, $\sigma_b = 0.002$

$\mathrm{RI}_{n,-3} = -0.225 + 0.025n$, $N = 51$, $r = 0.75$, $\sigma_e = 0.095$, $\sigma_a = 0.054$, $\sigma_b = 0.003$

$\mathrm{RI}_{n,+1} = -0.129 + 0.020n$, $N = 41$, $r = 0.68$, $\sigma_e = 0.099$, $\sigma_a = 0.060$, $\sigma_b = 0.003$

Allele frequency estimates in pooled DNA samples: *Constructed pools:* The frequency estimates from the constructed pools (means of four amplifications), plotted against the known frequencies, are shown in Figure 3. In all, pool estimates of allele frequencies were obtained for 60 bands. The correlation between the pool estimates and the known frequencies was high ($r = 0.88$). The estimates were somewhat biased, with the intercept and the coefficient of the regression equation $y = 0.04 + 0.83x$ differing significantly from 0.0 ($P < 0.05$) and 1.0 ($P < 0.01$).

Selective DNA pooling: Allele repeat numbers were estimated on the basis of measured $\mathrm{RI}_{n,-1}$ (data not shown). Densitometric estimates of allele frequencies were obtained for a total of 58 bands. The correlation between the pool estimates and the individual estimates was high, $r = 0.94$ (Figure 3). The regression equation was $y = 0.01 + 0.89x$. The intercept did not differ significantly from 0.0, while the regression coefficient again differed significantly from 1.0 ($P < 0.05$). Close inspection of Figure 3, however, revealed clusters of positive frequency estimates in the pools at zero frequency for individual genotyping. The positive bias is mainly an artifact of the calculation method: Since there cannot be negative frequencies, all negative CI were zeroed before frequency calculations, while corresponding positive values were allowed to stand. In all, for the selective DNA pools, there were only five cases where the CI for a band was zero or negative, but a corresponding rare allele was found among the individual genotypes. There were, however, 16 cases where the CI of a band was positive, while the corresponding allele was not found on individual genotyping. Calculating the regression only for those alleles for which individual genotyping gave a minimum frequency of one allele in the pool gave a new regression equation, $y = 0.00 + 0.93x$ ($r = 0.93$), for which both the intercept and the regression coefficient did not differ significantly from 0.0 and 1.0. Errors in identification of rare alleles have no influence on the significance tests and on allele-substitution-effect calculations of the selective DNA pooling method, because these are based on the sire alleles (Darvasi and Soller 1994), which are always present at high frequencies. For population studies based on DNA pooling, however, it may be advisable to set a minimum threshold for inclusion of rare alleles in the final frequency estimates.

Testing for marker-QTL linkage by selective DNA pooling: The total number of listed daughters per sire ranged from 1827 to 3407. Consequently, the proportion targeted for selection over both tails (a total of 440 cows per sire) ranged from 0.12 to 0.24. Out of 110 target cows listed for each external or internal pool, 68–89 were actually sampled. Missing animals and minor errors during pool preparation caused only insignificant differences in mean BV $P\%$ in the actual pools as compared to the target pools (data not shown).

Out of 420 individually genotyped milk samples, 22 (5.2%) had more than two alleles. These may represent mixing of milk samples or embryonic blood mixing in twins. Forty daughters (9.5%) did not carry any alleles of their putative sire, and hence are instances of false parentage identification.

Technical variance, V_T, did not differ significantly between markers by Bartlett's test. For this reason we used the value of V_T pooled over all markers, $V_T = 0.000722$, as the estimate of V_T in all calculations.

The frequency of the sire allele among the dam populations, *i.e.*, M_L and M_S, ranged from 0.0 to 0.70, with one exceptional value of 0.95. The average M value was 0.17, giving an $M(1 - M)$ value of 0.14, and this value was used in subsequent theoretical power calculations. However, in calculating the significance of marker-QTL linkage, the actual individual M_L and M_S values found were used for each individual sire-marker combination.

The average pool included 81 daughters. Of these, 15% were taken to represent the above instances of mixing or false parentage identification, so that a standard value, $N = 70$, was used in all calculations of the binomal error component. Using this value for N and values for M_L and M_S as obtained for the individual sires,

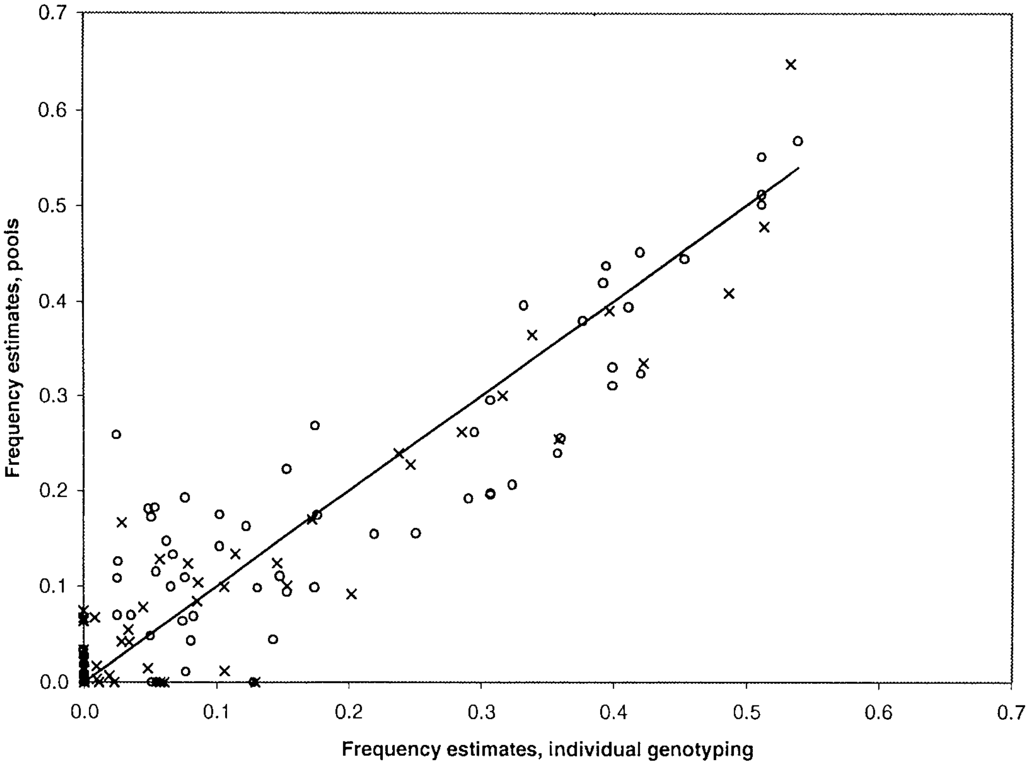

Figure 3.—Pool estimates. Comparisons of pool shadow-corrected estimates and those based on individual genotyping. —, $y = x$, expected regression line if pool estimates (y) are an unbiased estimate of the frequency based on individual genotyping (x); X, estimates based on constructed pools; ○, estimates based on pools of selective DNA genotyping study.

a first round of $SE^2(D)$ values was calculated on the assumption Cov $D_LD_S = 0$. On the basis of these values for $SE^2(D)$, six markers, CSSM19, BM6437, CSSM47, HUJI-74, RM6, and UWCA20, did not show significant association with BV $P\%$ ($P > 0.01$). The value of Cov D_LD_S calculated for these six markers was -0.00140. Using this value, $SE^2(D)$ values were recalculated, and a second round of significance tests performed. Marker ILSTS006 now lost significance. Including it among the nonsignificant markers increased Cov D_LD_S to -0.00185. Using this value, $SE^2(D)$ values were recalculated and a third round of significance tests performed. None of the remaining markers lost significance. Significance tests based on the final value of Cov D_LD_S are shown in Table 2. Even with this value of $SE^2(D)$, overall P value for marker ILSTS006, based on external pools, was less than 0.05, and one sire showed a P value of 0.002. This suggested that this marker may have been associated with a QTL affecting BV $P\%$. This was indeed shown to be the case when internal pools were analyzed. For this reason, the Cov D_LD_S value obtained without this marker, -0.0014, was used in subsequent theoretical power calculations.

There was an average of 4.6 heterozygous sires per marker, giving a total of 51 external-pool tests. Of these, 5 were significant at $P < 0.05$ and 9 were significant at $P < 0.01$. Thus, the number of significant tests was far greater than would have been expected by chance alone. The distribution of significant effects among the markers was highly uneven. Six of the markers, comprising 26 marker-sire combinations, did not have any significant test ($P < 0.05$) among them. None of these markers were significant overall.

Six markers were evaluated for both external and internal pools. Five of these markers were found significant. On the basis of combined pools, and pooled over all sires, P values of the five markers ranged from 3×10^{-4} to 3×10^{-16}.

Using the values $M = 0.17$, $N = 70$, and Cov $D_LD_S = -0.00185$, an overall calculated value of $SE(D_E) = 0.052$ is obtained. This was virtually the same as the empirical $SE(D_E)$, based on all DE values (L and S) for the six nonsignificant markers.

Allele substitution effects: Estimates of allele substitution effects, α_T, for individual sire-marker analyses, based on significant D_C values only, were 0.014 for INRA003, sire 2; 0.011 and -0.013 for RM188, sires 1 and 7, respectively; -0.016, -0.019, -0.024, and -0.032 for BM143, sires 2, 3, 4, and 7, respectively; -0.016, -0.016, and -0.021 for CSN3, sires 3, 4, and 5, respectively; and -0.022 for ILSTS006, sire 3. Mean substitution effects averaged over sires ranged from 0.012 to 0.023 BV $P\%$.

Reliability of pool estimates: Allele substitution effects were also estimated from the individual genotyping results for the three marker-sire combinations for which such data were available. For this analysis all mixed samples and all samples indicating false parentage identification were removed from the data. Estimated effects based on external pools or individual genotyping were, respectively: 0.016 and 0.010 for sire 2, marker INRA003; -0.032 and -0.030 for sire 7, marker BM143; and

TABLE 2

Marker-QTL linkage, according to chromosome, markers, and sires

Chromosome markers[a]			Sires							
			1	2	3	4	5	6	7	P
1	CSSM19	D_E	0.10*	0.06	−0.03	—	−0.03	0.01	0.07	2×10^{-1}
3	INRA003	D_E	—	0.15***[b]	—	−0.01	—	0.11**	—	
		D_I	—	0.14***	—	−0.02	—	0.03	—	
		D_C	—	0.14****	—	0.01	—	0.07*	—	8×10^{-5}
3	BM6437	D_E	—	−0.01	—	—	0.02	0.02	—	5×10^{-1}
4	RM188	D_E	0.14***	—	−0.06	0.04	−0.13****	−0.01	−0.13**	
		D_I	0.05	—	−0.05	0.01	−0.10*	0.01	−0.08	
		D_C	0.09***	—	−0.05	0.03	−0.02	0.03	−0.11***	3×10^{-4}
6	BM143	D_E	−0.06	−0.17****	−0.13***	−0.11****	−0.05	−0.01	−0.20*[b]	
		D_I	0.04	−0.12***	−0.10**	−0.11***	0.00	−0.06	−0.20****	
		D_C	−0.01	−0.15****	−0.11****	−0.11****	−0.02	−0.04	−0.20****	3×10^{-16}
6	CSN3	D_E	—	—	−0.14***	−0.12[b]	−0.16***	0.14**	—	
		D_I	—	—	−0.11**	−0.13**	−0.14***	−0.01	—	
		D_C	—	—	−0.13****	−0.12****	−0.15****	−0.07*	—	1×10^{-8}
7	RM006	D_E	0.05	−0.02	0.09	0.01	—	—	—	4×10^{-1}
7	UWCA20	D_E	0.05	−0.03	0.09*	−0.03	—	−0.07	0.04	
		D_I	0.09*	−0.05	0.05	−0.03	—	−0.03	0.04	
		D_C	0.07****	−0.05	−0.02	−0.06*	—	−0.05	0.04	7×10^{-2}
7	ILSTS006	D_E	−0.06	—	−0.16***	−0.01	0.05	0.01	—	
		D_I	0.22****	—	−0.12**	−0.06	0.07	0.04	—	
		D_C	0.08	—	−0.14****	−0.03	0.06*	0.02	—	2×10^{-4}
8	HUJI-74	D_E	−0.01	—	0.04	—	0.07	—	—	5×10^{-1}
8	CSSM47	D_E	—	−0.00	−0.05	—	0.00	—	—	8×10^{-1}

D_E, D_I, D_C, differences in sire allele frequencies between highland low pools, based on external pools, internal pools, and combined pools, respectively; P, overall P value associated with the marker. (See text for details.)

*$P < 0.10$; **$P < 0.05$; ***$P < 0.01$; ****$P < 0.001$.

[a] Barendse *et al.* (1994); Bishop *et al.* (1994).

[b] Confirmed by individual genotyping.

−0.013 and −0.012 for sire 4, marker CSN3 (note that these values are based on the *external* pools and slightly different from the values of the *combined* pools reported in the former section). Thus, there was a good accord between the estimates of substitution effects based on individual genotyping and those based on pool densitometry.

For the six markers for which both external and internal pools were evaluated, it is instructive to compare D_E values based on external pools and D_I values based on internal pools, according to whether the effects in the combined pools, D_C, were significant. Of 10 significant D_C values, the correlation between D_E and D_I values was $r = 0.97$ ($P < 0.001$). Of the 20 nonsignificant D_C values, the correlation between values for external and internal pools was 0.07 ($P > 0.05$). The very strong correspondence between external and internal pools for significant tests, in contrast to the weak correspondence for nonsignificant tests, is as expected if the significant effects indeed represent marker-QTL linkage.

DISCUSSION

False parentage identification: The rate of false parentage identification found in this study (9.5%) is in accord with other results (Ron *et al.* 1996a) and with results from our laboratory (Lipkin 1996). Mixing of blood can happen naturally with twins (Lipkin *et al.* 1993b), and milk somatic cells are actually blood leukocytes, so that some part of the 5% of mixed samples may be due to this factor. These "erroneous" samples are completely hidden in the pools, while individual genotyping enables their elimination. Nevertheless, the consequent reduction of the effective pool size from $N = 80$ to 70 had only a negligible influence on the results of the significance tests (data not shown), and there was a close correspondence between allele substitution effect estimates based on pools and those based on individual genotyping, among which these samples were not included.

Marker-QTL linkage: Five markers out of the 11 examined had an overall significance level in the range $P < 0.001$, which is considered significant for multiple marker tests (Lander and Kruglyak 1995). These markers also showed good correspondence between external and internal pools and multiple sire-within-marker significant effects. On this basis they were considered to represent marker-QTL linkages. The high proportion of markers showing significant effects may be due to the fact that 4 of these 5 markers were preselected as candidates for such linkage on the basis of the literature review. Within these 5 markers, 12 of 25 individual sire-by-marker tests were confirmed as significant by combined pools. This high proportion of individual sires showing significant effects is a result of two factors. The first is the high statistical power of the very large half-sib families of the A.I. sires (Soller 1990). Essentially, this high power means that any sire that is heterozygous at a QTL having an appreciable effect will show a significant effect for a marker in close linkage to that QTL. The second factor is that even for a diallelic QTL, heterozygosity will be close to 0.5 over a broad range of allele frequencies (from 0.3 to 0.7 for the positive allele). For a high heritability trait, such as milk protein percentage, it is plausible that some QTL are in the intermediate frequency range. It is these QTL that produce most of the significant marker-associated effects.

Four of the five markers showing linkage to QTL confirmed previous studies. In particular, INRA003 is 3 cM from TGLA263 (Barendse *et al.* 1994), which was found linked to milk protein and fat percentage by Ron *et al.* (1996b), although not by George *et al.* (1995). RM188 is 35 cM from TGLA116 (Barendse *et al.* 1994), which was found linked to Weaver disease and loosely linked to milk production (George *et al.* 1993; Medjugorac *et al.* 1996). Linkage was not found for BM6437, located closer to the other side of TGLA116 from RM188 (Barendse *et al.* 1994; Bishop *et al.* 1994). This implies that the relevant QTL may be located between RM188 and TGLA116. CSN3, within the casein gene cluster, was found to be linked to milk production and fat percentage and kilograms of fat (Velmala *et al.* 1995).

BM143 is on the same chromosome and about 31 cM from CSN3 (Barendse *et al.* 1994; Bishop *et al.* 1994). This marker was identified as linked to milk protein percentage by Spelman *et al.* (1996). In addition, TGLA37, about 9 cM from BM143 and 22 cM from CSN3 (Barendse *et al.* 1994), was linked to milk production and percentage of protein and fat (George *et al.* 1995; Kühn *et al.* 1996), and ILSTS97, about 18 cM from BM143 and about 13 cM from CSN3 (Barendse *et al.* 1994), was linked to milk production (Kühn *et al.* 1996). We note that the effect on milk protein percentage found by George *et al.* (1995) as linked to TGLA37 (0.011) is half the mean substitution effect found here with BM143, and both effects differ greatly from the mean substitution effect of 0.123 found linked to BM143 by Spelman *et al.* (1996).

Of four sires heterozygous at both of the chromosome *6* markers used in this study (Table 2), all showed significance at $P < 0.1$ to $P < 0.001$ with CSN3, while two (sires 5 and 6) were far from significance with BM143. This implies that the two markers may be linked to different QTL.

ILSTS006 is the only marker showing linkage to QTL affecting milk protein percentage with no previous report on a linkage to milk and/or milk components. This marker is about 10 cM from the gene RASA and the marker AGLA260 for which George *et al.* (1995) did not find any linkage to milk production or milk composition.

The repeatability of the linkage tests in different studies means that at least some QTL effects are robust, coming to expression in different environments. As such, they provide a firm base for marker-assisted selection (MAS) programs (Soller 1994).

Effective mapping population size on alternative genotyping schemes: It is instructive to compare the effective population size for marker-QTL mapping through total population genotyping, selective genotyping, and selective DNA pooling, according to the parameters obtained in this study. For the seven sires in this study, there were a total of 17,754 milk-recorded daughters listed as available for sampling at the time the milk samples were collected. Of these, 440 per sire, or a total of 3080, were targeted for sampling. In the event, a total of 2270 samples (73.7%) were actually collected. Thus, the listed daughters represented an actual available population of 13,085 daughters, or an average of 1869 per sire. Of these, the extreme high and low 17.3% were actually sampled, *i.e.*, 324 daughters per sire.

Following Darvasi and Soller (1992, expressions 18 and 25), the 2270 sampled daughters are the equivalent of 7899 daughters on total genotyping (1128 per sire). That is, selective genotyping of 17.3% of the total population would have retained 60% (7899/13,085) of the effective statistical power of the entire population. Note that after deducting 10% of the daughters as parentage errors, and an additional 5% as mixed samples, the 13,085 available daughters carry the statistical power of 11,122 daughters. From this should be deducted a further 5% representing daughters that are heterozygous for the same pair of marker alleles as their sire. This leaves an effective number of 10,565 daughters for total population genotyping and 60% of this, or 6340 daughters for selective genotyping.

To calculate the equivalent number for selective DNA pooling, we will first follow Darvasi and Soller (1994, expression 5 noting that $\delta = d/2$) to calculate effective type II error, β, for selective DNA pooling, according to the parameters obtained in the present study and assuming type I error, $\alpha = 0.01$, and a codominant QTL ($h = 0$) with standardized allele substitution effect, $d = 0.2$. We then substitute these values for α, β, and d in Soller *et al.* (1976), to obtain an estimate of N', the equivalent sample size represented by a single genotypic group of a single sire. Multiplying this by 14 (two genotypic groups per sire, and seven sires) gives the effective population size attained by selective DNA pooling in the present study.

For the present study we have total proportion selected over both tails, $P = 0.173$. Also, we have the following: the effective average number of daughters per pool (after adjusting for mixing and parentage errors), $N = 70$; the average sire allele frequency in the dam population, $M = 0.17$; the technical error variance $V_T = 0.000722$; and Cov $D_L D_S = -0.0014$. From this we obtain SE(D_C) = 0.037. Then by Darvasi and Soller (1994, expression 5) we have $Z_\beta = 0.128$, and by Soller *et al.* (1976) we have $N' = 365.3$. Hence, the effective population size attained by selective DNA pooling was 5110. This is 80.6% of the effective population size provided by individual selective genotyping (6340 daughters), and 46.3% of that provided by total population genotyping (10,565 daughters).

Thus, genotyping eight pools per sire in selective DNA pooling provided equivalent statistical power to actual genotyping of 903 daughters per sire in individual genotyping (13,085 total genotyped daughters × 0.483/7 sires) or of 910 daughters per sire in selective genotyping (13,085 × 0.60 × 0.806/7), over a 100-fold reduction. In practice, the reduction was even greater than this, since a sequential procedure was used such that most sire-marker combinations were evaluated on the basis of four pools (one external high and one external low pool, two replicates per pool), and only those that gave an indication of significance were evaluated on the four additional internal pools. Consequently, in the present study 51 sire-marker combinations were evaluated on the basis of 328 genotyping runs (51 external and 31 internal pool pairs, two replicates of each pool) or 6.4 genotyping runs per combination. This provided statistical power equivalent to 45,600 individual genotypings (11 markers × 4.6 heterozygous sires per marker × 903 daughters per sire), a 140-fold reduction. It should be noted, however, that individual genotyping provides information for mapping all traits, while selective DNA pooling is applied to one trait at a time. Thus, the reduction for a six-trait study, say, would be only 25-fold.

QTL phase identification for marker-assisted selection: When markers and QTL are in linkage equilibrium, schemes for MAS based on marker-QTL linkage (Kashi *et al.* 1990) require determining the specific coupling relations between marker alleles and QTL alleles in each elite proven sire for each located QTL. Determining marker-QTL coupling for an individual elite sire will require individual genotyping of some 1000 daughters of the sire with respect to 20 to 30 markers of interest, a total of 20,000 to 30,000 individual genotyping runs. Individual selective genotyping will not be useful here, since a number of different traits are involved. Using DNA pooling for the six traits of major importance, the sire can be evaluated on the basis of eight pools per marker, or 160 to 240 runs all told, a reduction of at least 125-fold as compared to individual genotyping. An onerous and costly running expense for MAS scheme is thus converted into a virtual triviality. In this case, the full benefit of selective DNA pooling is obtained, even in a multitrait situation.

QTL mapping of the bovine genome using selective DNA pooling: In the present study a total of 2270 milk samples were collected from among the daughters of seven sires in active artificial insemination (A.I.) service in Israel. The samples were collected during routine milk recording in commercial Israeli herds. The entire sampling effort took less than 2 months, at a cost of about $0.50 per sample. Pool preparation was a matter of days, at practically no additional cost. The sampled daughters provided statistical power equivalent to that given by 731 effective daughters per sire (after adjusting for mixing and parentage errors) with total individual genotyping. Table 1 of Weller *et al.* (1990) with five sires and 800 daughters per sire confirms the very high overall power for selective DNA pooling as expressed in the present study: Power is 0.76 for an allele substitution effect of $\alpha = 0.2$ and 0.94 for $\alpha = 0.3$. This was achieved at an average of 6.4 genotyping runs per sire-marker combination. Extending this to 100 markers and the six traits of major importance to dairy cattle breeding (yearly kilograms of milk production, protein and fat percentage, milk somatic cell count, female fertility, and calf mortality) implies that total genome mapping at high power could be achieved with a total of 15,500 genotyping runs. This should bring total dairy cattle genome mapping within reach of most AI centers.

Fine mapping by selective DNA pooling: Darvasi (1997) has shown that the confidence interval of QTL map location with selective genotyping stands in the same proportion to total genotyping as does power. Thus, selective DNA pooling can potentially harness the very large dairy cattle half-sib sire families for purposes of fine mapping. Darvasi and Soller (1997) have shown that the 95% confidence interval of QTL mapping using a saturated genetic map stands in inverse linear proportion to sample size, N, according to the expression $3000/Nd^2$ for a backcross design. The daughter design used in the present experiments essentially represents a backcross design for each individual sire. Thus, for each sire, with the above effective $N = 731$, we have a confidence interval of 103 cM for an allele substitution effect $\alpha = 0.2$, and 46 cM for an effect of $\alpha = 0.3$. However, mapping results can be accumulated over sires. As a result, a set of 10 sires heterozygous at a QTL would provide a 95% confidence interval of 10.3 cM for a substitution effect $\alpha = 0.2$, and 4.6 cM for a substitution effect of $\alpha = 0.3$. This set of sires could be accessed by a single A.I. center over the course of a few years. If results are combined across A.I. centers, 30 sires heterozygous for a given QTL would provide confidence intervals of 3.4 and 1.5 cM for QTL of effects $d = 0.2$ and 0.3, respectively. Identifying suitable candidate genes in regions of this size through synteny of human, mouse, and bovine gene maps (Womack and Kata 1995) should be eminently possible. Assuming that the pools are already available as part of the routine QTL mapping and sire evaluations, the total genotyping costs of such a fine mapping project would be negligible.

The methods described here can reduce QTL mapping costs to the point where an organization of moderate size can do its own total genome mapping. This may reduce the marginal commercial value of proprietary information below the commercial value of the fine mapping information that can be obtained by sharing sample and data bases.

We thank Alona Nave, Einat Azarya, and Noa Sagiv for dedicated technical assistance and the Israeli Cattle Breeders Organization for the use of the herd book and for sampling the milk and conducting the cell count. This study was supported by the joint Biotechnology program of the Ministry of Science and Technology (Israel) and Bundesministerium für Biotechnologie (Germany), by the Israel Science foundation, and by the Dairy Council (Israel). M.O.M was supported by the German Academic Exchange Service (DAAD).

LITERATURE CITED

Barendse, W., S. M. Armitage, L. Kossarek, A. Shalom, B. Kirkpatrick *et al.*, 1994 A genetic linkage map of the bovine genome. Nature Genetics **6:** 227–234.

Bishop, M. D., S. M. Kappen, J. W. Keele, R. T. Stone, S. L. F. Sunden *et al.*, 1994 A genetic linkage map for cattle. Genetics **136:** 619–639.

Cheng H., I. Levin, L. Vallejo, H. Khatib, J. B. Dodgson *et al.*, 1995 Development of a genetic map of the chicken with markers of high utility. Poult. Sci. **74:** 1855–1874.

Crawford, A. M., K. G. Dodds, A. J. Ede, C. A. Pierson, G. Montgomery *et al.*, 1995 An autosomal genetic linkage map of the sheep genome. Genetics. **140:** 703–724.

Darvasi, A., 1997 The effect of selective genotyping on QTL mapping accuracy. Mamm. Genome **8:** 67–68.

Darvasi, A., and M. Soller, 1992 Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet. **85:** 353–359.

Darvasi, A., and M. Soller, 1994 Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics **138:** 1365–1373.

Darvasi, A., and M. Soller, 1997 A simple method to calculate resolving power and cofidence interval of QTL map location. Behav. Genet. **27:** 125–132.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot *et al.*, 1996 Comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

Dietrich, W., F. J. Miller, R. Steen, M. A. Merchant, D. Damron-Boles *et al.*, 1996 A comprehensive genetic map of the mouse genome. Nature **380:** 149–152.

George, M., A. B. Dietz, A. Mishra, D. Nielsen, L. S. Sargeant *et al.*, 1993 Microsatellite mapping of the gene causing weaver disease in cattle will allow the study of an associated quantitative trait locus. Proc. Natl. Acad. Sci. USA **91:** 1058–1062.

George, M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto *et al.*, 1995 Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics **139:** 907–920.

Hauge, X. Y., and M. Litt, 1993 A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. Hum. Mol. Genet. **2:** 411–415.

Kashi Y., E. Hallerman and M. Soller, 1990 Marker-assisted selection of candidate bulls for progeny testing programs. Anim. Prod. **51:** 63–74.

Khatib, H., A. Darvasi, Y. Plotsky and M. Soller, 1994 Determining relative microsatellite allele frequencies in pooled DNA samples. PCR Methods Appl. **4:** 13–18.

Kühn, C. H., R. Weikard, G. Freyer and M. Schwerin, 1996 Detection of QTL for milk production traits on chromosome *6* in German Holstein Friesian cattle. Anim. Genetics **27**(Suppl.): 103.

Lagziel, A., E. Lipkin and M. Soller, 1996 Association between SCCP haplotypes at the bovine growth hormone gene and milk protein percentage. Genetics **142:** 945–951.

Lander E. S., and L. Kruglyak, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature Genet. **11:** 241–247.

Leduc, C., P. Miller and P. Parry, 1995 Batched analysis of genotypes. PCR Methods Appl. **4:** 331–336.

Levinson, G., and G. A. Gutman, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

Lipkin, E., A. Shalom, H. Khatib, M. Soller and A. Friedmann, 1993a Milk as a source of deoxyribonucleic acid and as a substrate for the polymerase chain reaction. J. Dairy Sci. **76:** 2025–2032.

Lipkin, E., Y. Tikoschinsky, R. Arbel, D. Sharoni, M. Soller *et al.*, 1993b Early PCR amplification test for identifying chimerism in female calves co-twin to a male in cattle. Anim. Biotech. **4:** 195–201.

Lipkin, E., 1996 Use of pooled DNA samples for the estimation of microsatellites allele frequencies by direct PCR from milk samples. Ph.D. Thesis (in Hebrew with English summary), Hebrew University, Jerusalem.

Litt, M., C. Hauge and V. Sharma, 1993 Shadow bands seen when typing polymorphic dinucleotide repeats: some causes and cures. BioTechnology **15:** 280–284.

Medjugorac, I., I. Russ, J. Aumann and M. Förster, 1996 Weaver carrier status effects on yield in German Brown cattle. Anim. Genet. **27:** 105.

Murray, V., C. Monchawin and P. R. England, 1993 The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. Nucl. Acids Res. **21:** 2395–5398.

Ostensson, K., M. Hageltorn and G. Astrom, 1988 Differential cell counting in fraction-collected milk from dairy cows. Acta Vet. Scand. **29:** 493–500.

Perlin, M. W., G. Lancia and S. K. Ng, 1995 Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. Am. J. Hum. Genet. **57:** 1199–210.

Ron, M., Y. Blanc, M. Band, E. Ezra and J. I. Weller, 1996a Misidentification rate in the Israeli dairy cattle population and its implication for genetic improvement. J. Dairy Sci. **79:** 676–681.

Ron, M., D. W. Heyen, M. Band, E. Feldmesser, Y. Da *et al.*, 1996b Detection of individual loci affecting economic traits in the US Holstein population with the aid of DNA microsatellites. Anim. Genet. **27**(Suppl.): 105.

Shalom, A., M. Soller and A. Friedmann, 1993 Dinucleotide repeat polymorphism at the bovine locus HUJ616. Anim. Genet. **24:** 327.

Shalom, A., M. O. Mosig, W. Barendse, A. Friedmann and M. Soller, 1994 Dinucleotide repeat polymorphism at the bovine HUJII77, HUJ223, HUJVI174, HUJ175 loci. Anim. Genet. **25**: 56.

Soller, M., 1990 Genetic mapping of the bovine genome using DNA-level markers with particular attention to loci affecting quantitative traits of economic importance. J. Dairy Sci. **73:** 2628–2646.

Soller, M., 1994 Marker assisted selection—an overview. Anim. Biotech. **5:** 193–207.

Soller, M., T. Brody and A. Genizi, 1976 On the power of experimental designs for detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. **47:** 35–39.

Spelman R. J., W. Coppieters, L. Karim, J. A. M. van Arendock and H. Bovenhuis, 1996 Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. Genetics **144:** 1799–1808.

Tautz, D., 1989 Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucl. Acids Res. **17:** 6463–6471.

Velmala, R., J. Vilkki, K. Elo and A. Maki-Tanila, 1995 Casein haplotypes and their association with milk production traits in the Finnish Ayrshire cattle. Anim. Genet. **26:** 419–425.

Walpole, R. E., and R. H. Myers, 1972 *Probability and Statistics for Engineers and Scientists*, p. 375. Macmillan Publ. Co. Inc., NY.

Weber, J. L., and P. E. May, 1989 Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. **44:** 388–396.

Weller, J. I., Y. Kashi and M. Soller, 1990 Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. J. Dairy Sci. **73:** 2525–2537.

Womack, J. R., and S. R. Kata, 1995 Bovine genome mapping: evolutionary inference and the power of comparative genomics. Curr. Op. Genet. Devel. **5:** 725–733.

Communicating editor: Z-B. Zeng