# Fine-mapping milk production quantitative trait loci on BTA6: Analysis of the bovine osteopontin gene

**Robert D. Schnabel*†‡, Jong-Joo Kim§, Melissa S. Ashwell†, Tad S. Sonstegard†, Curtis P. Van Tassell†, Erin E. Connor†, and Jeremy F. Taylor***

*Division of Animal Sciences, University of Missouri, Columbia, MO 65211; †Bovine Functional Genomics Laboratory, U.S. Department of Agriculture/Agriculture Research Service, Beltsville, MD 20705; and §School of Biotechnology, Yeungnam University, Gyeongsan 712-749, Korea

Communicated by R. Michael Roberts, University of Missouri, Columbia, MO, March 23, 2005 (received for review February 11, 2005)

**Bovine chromosome six (BTA6) harbors up to six quantitative trait loci (QTL) influencing the milk production of dairy cattle. In stark contrast to human, there is long-range linkage disequilibrium in dairy cattle, which has previously made it difficult to identify the mutations underlying these QTL. Using 38 microsatellite markers in a pedigree of 3,147 Holstein bulls, we fine mapped regions of BTA6 that had previously been shown to harbor QTL. Next, we sequenced a 12.3-kb region harboring Osteopontin, a positional candidate for the statistically most significant of the identified QTL. Nine mutations were identified, and only genotypes for the *OPN3907* indel were concordant with the QTL genotypes of eight bulls that were established by segregation analysis. Four of these mutations were genotyped, and a joint linkage/linkage disequilibrium mapping analysis was used to demonstrate the existence of only two functionally distinct clusters of haplotypes within the QTL region, which were uniquely defined by *OPN3907* alleles. We estimate a probability of 0.40 that no other mutation within this region is concordant with the QTL genotypes of these eight bulls. Finally, we demonstrate that the motif harboring *OPN3907*, which is upstream of the promoter and within a region known to harbor tissue-specific osteopontin regulatory elements, is moderately conserved among mammals. The motif was not retrieved from database queries and may be a novel regulatory element.**

linkage disequilibrium

Since the first genome scan to detect quantitative trait loci (QTL) in dairy cattle (1), QTL affecting milk production traits have been identified on almost every bovine autosome (2). A QTL proximal to the centromere of BTA14 that affects milk fat percent (FP) has consistently been identified. The causal mutation underlying this QTL has independently been identified by two groups (3, 4) as a K232A substitution in exon VIII of *DGAT1* (acylCoA/diacylglycerol acyltransferase 1). Most studies have also identified QTL on BTA6, consistently identifying a QTL affecting milk protein percent (PP) near *BM143* (5). The genes and causal mutations underlying the BTA6 milk QTL have yet to be identified. However, several recent reports have focused upon the PP QTL near *BM143*. This QTL was localized to a 4-cM region around *BM143* (55.4 cM) in an Israeli Holstein population where two additional QTL near *BM415* (80.5 cM) and the centromere were also identified (6). Freyer *et al.* (7) reported two QTL for milk yield (MY) at 41 and 91 cM and two QTL for PP at 44 and 67 cM, as well as a QTL affecting both fat yield (FY) and protein yield (PY) at 70 cM. Olsen *et al.* (8) localized the FP and PP QTL near *BM143* to a 7.5-cM interval bounded by *BMS2508* and *FBN12*. Cohen *et al.* (9) were the first to begin sequencing candidate genes in this region and, whereas *FAM13A1* appeared to be a likely functional candidate, it was excluded as underlying the QTL, which was placed centromeric of *FAM13A1*. This QTL has now been fine mapped to a 420-kb interval between genes *ABCG2* and *LAP3* (10). Within this region, there are only four known human genes: *IBSP*, *MEPE*, *OPN*, and *PKD2* (http://genome.ucsc.edu).

Osteopontin (*OPN*, *SPP1*, and *Eta-1*) is a strong functional candidate for this QTL. *OPN* is a secreted glycoprotein that functions by mediating cell–matrix interactions and cellular signaling through binding with integrin and CD44 receptors and is expressed in numerous tissues (11). Expression of *OPN* in the murine mammary gland depends on stage of postnatal development, suggesting a role for *OPN* in mammary involution (12). A transgenic mouse model expressing *OPN* antisense RNA in the mammary gland has shown *OPN* to be necessary for normal mammary gland development and lactation (13). *OPN*-antisense mice lack alveolar structures, suffer a drastic reduction in the synthesis of milk proteins $\beta$-casein and whey acidic protein, and are lactation-deficient. Finally, *OPN* is up-regulated 8.8-fold in parous involuted vs. nulliparous mammary glands of both mice and rats (14).

There is an increasing body of evidence that indicates BTA6 harbors the largest number of milk production QTL of any bovine chromosome. One of these QTL has been localized to a small critical region, but the underlying gene and causal mutation have yet to be identified. Here we present evidence for a putative quantitative trait nucleotide in the upstream regulatory region of bovine *OPN* with significant effects on MY, FP, and PP.

## Materials and Methods

**Animals and Traits.** DNA samples from Holstein sires were obtained from the Cooperative Dairy DNA Repository for 45 half-sib families (15). Each of these families belongs to one of three extended superfamilies denoted L, M, and N, which comprise three, five, and three generations of extended half-sib families, respectively (Table 2, which is published as supporting information on the PNAS web site). All three founding sires (L-0, M-I-1, and N-0) and all intermediary sires linking families to the founding sires were genotyped. Predicted transmitting abilities (PTAs) were obtained from the U.S. Department of Agriculture Animal Improvement Programs Laboratory (May 2004 evaluations). Traits analyzed were MY, FY, PY, FP, and PP.

**Genotyping.** Microsatellite markers ($n$ = 38; Table 3, which is published as supporting information on the PNAS web site) were scored in multiplex reactions by using established protocols (16). PCR products were separated on an Applied Biosystems 3700 and genotypes assigned by using GENOTYPER 3.7 (Applied Biosystems). All families were genotyped for the *DGAT1* K232A mutation (3). A BTA6 linkage (LK) map was constructed by using CRI-MAP, Ver. 2.4 (17), and GENOPROB (18, 19) was used to quality assure genotypes.

Abbreviations: QTL, quantitative trait locus; IBD, identity by descent; PTA, predicted transmitting ability, BF, Bayes factor; lod, logarithm of odds; MY, milk yield; FY, fat yield; PY, protein yield; FP, fat percent; PP, protein percent; LK, linkage; LD, linkage disequilibrium; *OPN*, osteopontin.

Data deposition: The sequence data reported in this paper have been deposited in the GenBank database (accession no. AY878328).

‡To whom correspondence should be addressed. E-mail: schnabelr@missouri.edu.

**QTL Analysis.** Three complementary approaches were used: (*i*) half-sib family analysis using QTL EXPRESS (20); (*ii*) full pedigree analysis using LOKI, Ver. 2.4.5 (21); and (*iii*) combined LK/LD (LD, linkage disequilibrium) analysis using LDVCM (22). All analyses used our male-specific genetic map in Haldane cM.

Each family was individually analyzed by using QTL EXPRESS to determine the sire's QTL segregation status for each trait. Data permutation ($n = 5{,}000$) was used to determine chromosome-wise significance levels for each sire (23). Tests of one vs. zero, one vs. two, and two vs. zero QTL were conducted individually for each family. Sires significant at the chromosome-wise $P < 0.05$ level for the one- or two-QTL models were classified as segregating, regardless of trait or QTL position. Data for all 26 segregating sires were combined into a data set. Across-family analysis was performed on this data set, and bootstrapping ($n = 1{,}000$) was used to obtain confidence intervals for QTL location (24).

Determination of significance levels by data permutation is not possible using the two-QTL model in QTL EXPRESS. To account for multiple testing in the two-QTL analyses, we used the following approach. For the one-QTL model, $F$ statistics were generated based on data permutation to represent the chromosome-wise $P < 0.05$ and $P < 0.01$ levels for each sire/trait combination. For example, the chromosome-wise $P < 0.05$ level based on data permutation for sire L-I-1 with 93 sons required an $F$ statistic of 6.33. The nominal $P$ value corresponding to $F = 6.33$ as an observation on an $F$ distribution with 1-numerator and 92-denominator df is $P = 0.0136$. Thus, for the two-QTL model for this sire to be considered significant at chromosome-wise $P < 0.05$, the uncorrected $P$ value associated with the two-QTL $F$ statistic must be $<0.0136$. Sires significant for the one-QTL model were evaluated for the two- vs. one-QTL models, and sires not significant for the one-QTL model were evaluated for the two- vs. zero-QTL models.

LOKI was used for multipoint QTL analysis by using the segregating sire data set and also to individually analyze each half-sib family to estimate both the number and positions of QTL. An initial burn-in of 1,000 iterations was followed by 500,000 iterations, with parameter estimates collected at each iterate. The model for trait $y$ ($n\times 1$; $n$ animals each with a single observation) is: $y = \mu + X\beta + \Sigma_{i=1}^{k} Q_i\alpha_i + Zu + e$, where $\mu$ is the overall trait mean; $\beta$ is an ($m\times 1$) vector of fixed effects and covariates; $\alpha_i$ is a ($2 \times 1$) vector of allele substitution effects for the $i$th biallelic QTL; $u$ is an ($n\times 1$) vector of random normally distributed additive residual polygene effects; $e$ is an ($n\times 1$) vector of normally distributed residuals; $k$ is the number of QTL in the model; and $X$ ($n \times m$), $Q_i$ ($n\times 2$), and $Z$ ($n \times n$) are known incidence matrices for the fixed, QTL, and polygenic effects. *DGAT1* genotypes were included in the model as a fixed effect. LOKI treats the number of QTL in the model as a random variable to be estimated and allows for unlinked QTL within the genome. We set the total genome length to 2,900 cM to fit unlinked QTL.

The LDVCM software has been described (22). First, the LK phases of all sires and sons were estimated (25). Then, to jointly exploit LD (female meioses) and LK (male meioses) information, the following linear model was fit at a putative QTL location ($p$): $y = \mu + X\beta + Z_h h + Zu + e$, where $Z_h$ ($n \times n_h$) is an incidence matrix relating maternal haplotypes (MH) of sons ($n_{\text{mat}}$) and sire haplotypes (SH, $2n_{\text{sire}}$) to individual sons ($n_h = 2n_{\text{sire}} + n_{\text{mat}}$), and $h$ ($n_h \times 1$) is the vector of random haplotype effects (22). The remaining effects are as defined for LOKI. At most, three row elements of $Z_h$ are nonzero: a 1 in the column corresponding to the MH of the son, and $\lambda_p$ and $1-\lambda_p$ (probabilities that the son inherited the "left" or "right" SH at position $p$) in the columns corresponding to the two SH. Identity-by-descent (IBD) probabilities ($\varphi_p$) at the midpoint of each marker interval (defined as position $p$) were computed for all pairs of $n_h$ haplotypes conditional on the identity-by-state status of flanking markers (26). For the IBD computation, windows of 26 markers were considered as a haplotype, and the number of generations since the founder generation and the effective population size were both set to 100. A dendrogram was generated by using the unweighted pair group method with arithmatic mean (UPGMA) hierarchical clustering algorithm with $1-\varphi_p$ as the distance measure at QTL location $p$. Starting at the ancestral node and sequentially descending into the dendrogram, all possible combinations of haplotype clusters were analyzed in place of individual haplotypes. This process identified the set of nodes at which the likelihood of the data were maximized (22). For this set, all haplotypes within a cluster are predicted to be functionally equivalent, whereas haplotypes in at least one different cluster are functionally different. We used AIREML (27) to predict haplotype cluster effects and estimate variance components for haplotype clusters ($\sigma_H^2$), polygenic effects ($\sigma_A^2$), and error terms ($\sigma_E^2$) at each nodal step into the dendrogram defined at location $p$. This procedure was repeated for each intermarker interval until maxima corresponding to QTLs were identified. This provides the maximum (residual) likelihood estimate of QTL location and haplotype clusters harboring functionally equivalent QTL alleles. The proportion of variance explained by the haplotype clusters was estimated by $2\sigma_H^2/(2\sigma_H^2 + \sigma_A^2 + \sigma_E^2)$. A logarithm of odds (lod) score was obtained under the QTL and no-QTL models ($y = \mu + X\beta + Zu + e$). Alternative analyses were also performed: (*i*) exploiting only LK information by ignoring all MH and regarding SH as distinct haplotype clusters (22) and (*ii*) marker association tests in which marker alleles were treated as haplotypes (28). lod score thresholds were determined at point-wise $P$ levels by considering $2 \times \ln(10) \times$ lod to be distributed as $\chi_1^2$ (models differ by one parameter; $\sigma_H^2$), because the QTL had previously been detected (29).

**Sequencing *OPN*.** PCR primers were developed within *OPN* exons to amplify and sequence flanking introns based on the TIGR consensus sequence TC152671 (www.tigr.org/tdb/tgi). After the introns were sequenced, primers were designed within flanking introns to sequence each exon. To sequence the 5′ and 3′ regions of the gene, bacterial artificial chromosome (BAC) 263K19 was identified from the CHORI-240 bovine BAC library by using overgo hybridization. Sequencing primers were used to obtain ≈5 kb of sequence upstream of the transcription initiation site and 200 bp past the polyA signal from this clone. From this sequence, PCR primers were developed to allow the complete sequencing of *OPN* and flanking regions. A total of eight sires were sequenced for the entire 12.3-kb region, four sires identified as segregating (*Qq*) for the QTL within the 420-kb critical region and four nonsegregating sires (*QQ* or *qq*). All primers are available from the authors.

**Power Calculations.** We define $P_0$ to be the probability there are no other polymorphisms within a chromosomal interval harboring a QTL that are concordant with the QTL genotypes of $N_{\text{HET}}$ segregating and $N_{\text{HOM}}$ nonsegregating individuals. Let $\lambda$ be the expected number of polymorphisms (SNPs) within the critical region that is approximately $\Delta\times\phi$, where $\Delta$ is the average SNP density, and $\phi$ is the physical size of the region. Also let $N_S$ represent the actual number of SNPs within the QTL region and $P_i$ be the allele frequency of the $i$th SNP. If the individuals are not closely related, their genotypes are independent, and $P_0$ is:

$$P_0 = P(N_s = 0) + \sum_{j=1}^{\infty} P(N_s = j) \cdot \prod_{i=1}^{j}\{1 - [2P_i(1 - P_i]^{N_{\text{HET}}}[1 - 2P_i(1 - P_i)]^{N_{\text{HOM}}}\}.$$

We assume a Poisson distribution for $N_S$ and approximate the probability that the $i$th SNP is not concordant with the known QTL genotypes with its expectation:

$$P_0 = \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} E[1 - [2P_i(1 - P_i)]^{N_{HET}}[1 - 2P_i(1 - P_i)]^{N_{HOM}}]^j.$$

For given values of $N_{HET}$ and $N_{HOM}$, we estimate the expectation by sampling $P_i$ from a $U$ (0, 1) distribution.

## Results

**QTL Analysis by QTL Express and LOKI.** A total of 3,147 individuals from 45 families were available. Twenty-six sires representing all three superfamilies were segregating for QTL affecting one or more of the five milk production traits (Table 4, which is published as supporting information on the PNAS web site). Eleven sires were significant for the two-QTL model. The across-family $F$ statistic profiles based on these 26 sires are in Fig. 5, which is published as supporting information on the PNAS web site. Peak test statistics in the across-family analyses were: MY at 59 and 67 cM, FP at 64 cM, PY at 61 cM, and PP at 64 cM. Because there are multiple QTL influencing milk traits on BTA6, the test statistic profiles for the one-QTL model do not provide robust estimates of the number or positions of segregating QTL. Similarly, the use of the bootstrap to estimate confidence intervals for QTL location under this model is inappropriate. However, examination of the distribution of the bootstrap replicates reveals clusters corresponding to locations that are consistent between traits; 0 cM (MY, FP, and PY), 59–61 cM (MY, PY and PP), and 64–68 cM (MY, FP and PP). Localization of QTL to these regions was supported by the individual-family analyses (Table 4).

Across-family results for PP, FP and MY using LOKI are in Fig. 1. LOKI indicated the presence of two QTL for FP at 57 cM [Bayes Factor (BF), 123] and 60 cM (BF, 88) and three QTL for PP at 59 cM (BF, 229), 89 cM (BF, 56) and 95 cM (BF, 86). The 95% highest posterior density interval for the PP QTL at 57 cM was 7.2 cM (55.0–62.2 cM). LOKI found no evidence for FY or PY QTL.

**QTL Analysis by LDVCM.** LDVCM was used to analyze the segregating sire data set. Analyses of PP and MY using LK information gave comparable results to those obtained in the QTL EXPRESS analyses (Fig. 2). The most likely QTL positions were 64 and 66 cM with lod = 9.3 and 2.2 (nominal $P$ = 6E-11; 0.0013). For PY, a QTL was detected at 79 cM with lod = 5.6 ($P$ = 4.2E-7), distal from the most likely position estimated by QTL EXPRESS (local maximum at 60 cM with lod = 4.5; data not shown). For FP and FY, the most likely QTL positions were at 63 cM with lod = 1.0 for both traits. When LD information was included, highly significant lod scores were obtained for FP, lod = 8.8 ($P$ = 1.9E-10); MY, lod = 6.4 ($P$ = 5.7E-8) and PP, lod = 23.4 ($P$ = 3.0E-25) at positions similar to those obtained in the LK analyses (Fig. 2). However, for FY the lod increased to 2.5 at 72 cM and for PY lod scores decreased to 4.1 at 85 cM and 1.5 at 60 cM (results not shown). Both the LOKI and LDVCM analyses indicate a second PP QTL near the casein cluster (88–90 cM) and a third near 95 cM. The locations of both QTL are consistent with estimates from the within-family analyses.

**Sequencing and Analysis of *OPN*.** *OPN* was identified as a functional candidate for the QTL affecting PP in interstitial BTA6. A total of 12.3 kb of *OPN* sequence was submitted to GenBank (accession no. AY878328). Nine SNPs were found in the eight sequenced sires. SNP locations are numbered according to position within AY878328, and the SNP haplotypes are presented in Table 5, which is published as supporting information on the PNAS web site. The four segregating sires chosen for sequencing (L-I-1, L-II-14, L-II-15, and M-II-9) all shared the PP-decreasing QTL allele IBD. Thus, it is likely that a single SNP is responsible for the detected variation in PP, and for any of the detected SNPs to be a candidate quantitative trait nucleotide, the SNP genotypes must be concordant with the QTL genotypes of all segregating and nonsegregating sires. The only concordant SNP was *OPN3907*, an indel located ≈1,240 bp upstream of the *OPN* transcription initiation site. This indel occurs within a polyT tract producing alleles of either 9 ($T_9$) or 10 ($T_{10}$) thymines and was scored as a fragment-length polymorphism from PCR amplicons generated with a fluorescently labeled primer. The *OPN3379*, *OPN3490*, and *OPN3492* SNPs were genotyped by allele-specific PCR (30). All four SNPs were genotyped in a panel of 167 sires that represent all of the lines within the Cooperative Dairy DNA Repository.

Because the bulls represented in this panel were born between 1952 and 1996, selection has resulted in a significant increase in

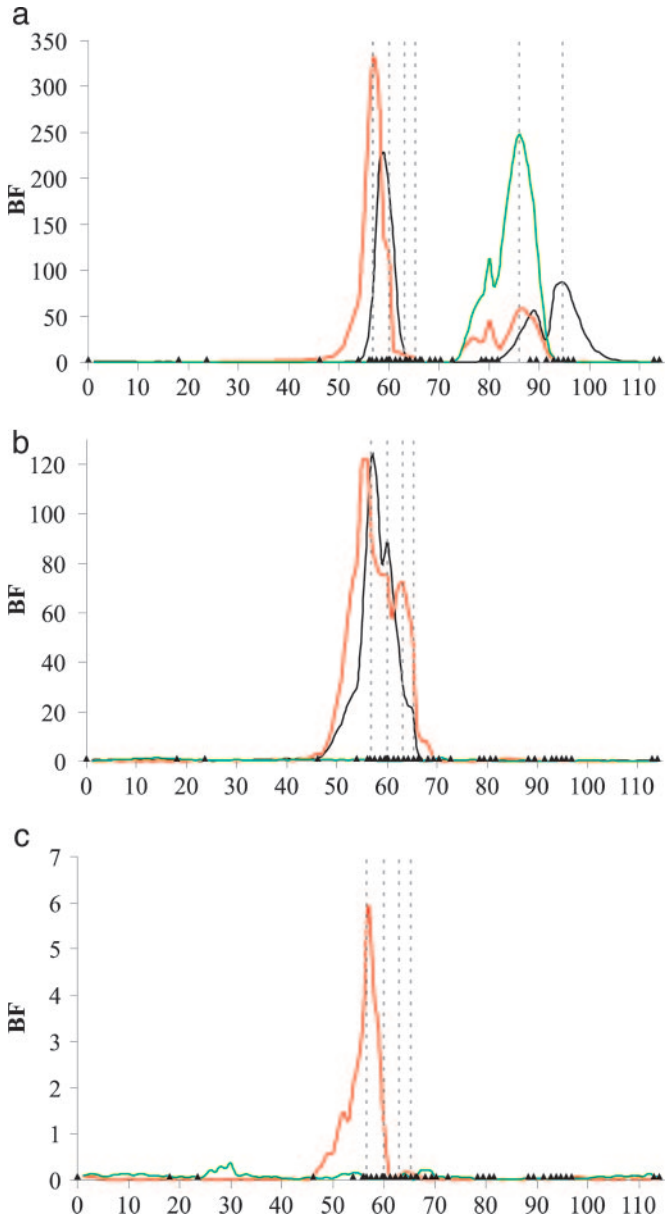


**Fig. 1.** BF profiles. PP (*a*), FP (*b*), and MY (*c*) using LOKI. The *x* axis is the chromosomal position (cM). Red (black) lines include (exclude) *OPN* SNPs in the map. Green lines fit *OPN3907* as a fixed effect. Gray dotted lines align peak locations between LOKI and LDVCM analyses.

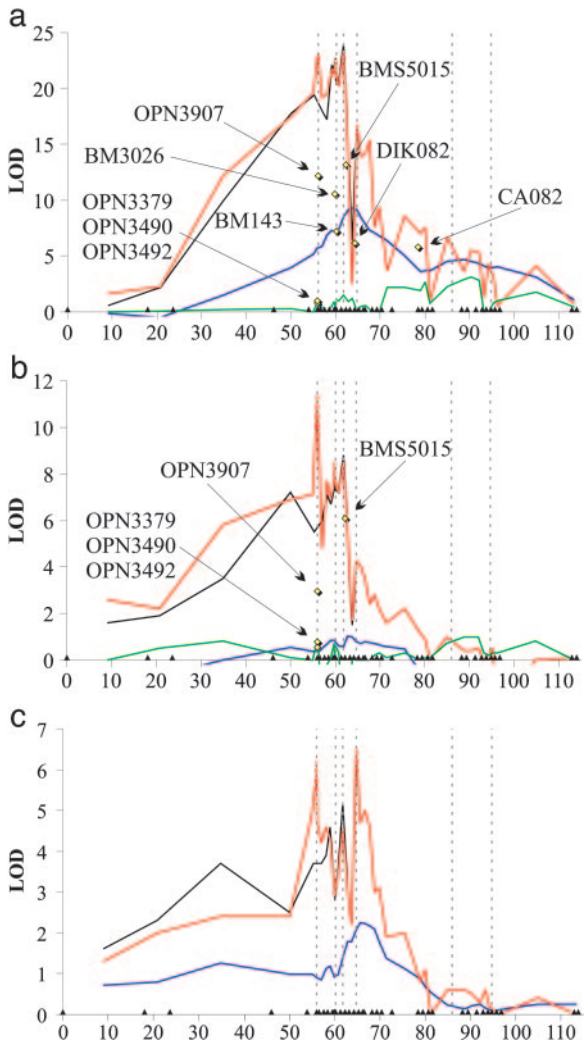


**Fig. 2.** lod profiles. PP (*a*), FP (*b*), and MY (*c*) using LDVCM. The *x* axis is the chromosomal position (cM). The blue line indicates LK, and remaining lines are combined LK/LD analyses. Red (black) lines include (exclude) *OPN* SNPs in the map. The green line fits *OPN3907* as a fixed effect. Yellow diamonds are individual marker analyses. Gray dotted lines align peak locations between LOKI and LDVCM analyses.

PTAs and changes in QTL allele frequencies. To eliminate biases due to genetic trend, we analyzed $M = \text{PTA}_{\text{bull}} - (\frac{1}{2}\text{PTA}_{\text{sire}} + \frac{1}{2}\text{PTA}_{\text{dam}})$, which is an estimate of one-half of the Mendelian sampling of parental gametes or one-half of the deviation of the mean value of the two gametes inherited by an individual from its parents from the average of all possible parental gametes. The variance of *M* will be larger in families that segregate for a major gene than in those not segregating, and *M* is independent of the rate of genetic trend in a population. The *M* values were analyzed by using ANOVA by contrasting animals with alternate genotypes at *OPN3907* (no $T_9/T_9$ genotypes were detected), *OPN3379*, *OPN3490*, and *OPN3492*. The only SNP with an effect on any trait was *OPN3907*, which influenced only PP ($P = 0.04$) (data not shown).

To better estimate the frequency and effect of *OPN3907*, 1,510 members of the superfamily *M* (Table 2) excluding families M-III-9 and M-III-12 were genotyped. Five families (M-II-1, M-III-10, M-IV-6, M-IV-8, and M-V-14) were also genotyped for *OPN3379*, *OPN3490*, and *OPN3492* to construct haplotypes and test the effects of these polymorphisms. All of the sires (except M-III-16) of these families were homozygous $T_{10}/T_{10}$ at *OPN3907*, and the $T_9$ alleles present in their progeny were maternally inherited, allowing an LD estimate of the effect of this SNP. The *M* values were again analyzed by ANOVA (Table 1). The *OPN3907* $T_9$ allele produced a 118.22-lb increase in MY ($P = 0.014$), a 3.98-lb decrease in FY (not significant), a 2.06-lb decrease in PY (not significant), a 0.0354% decrease in FP ($P = 1.36\text{E-}6$), and a 0.0242% decrease in PP ($P = 6.62\text{E-}14$).

*OPN3490* was significant for PP ($P = 0.005$, data not shown) but was excluded as the causal quantitative trait nucleotide, because sires L-I-1, L-II-14, and M-III-9 segregating for the QTL were all homozygous. The association appears to be due to LD, because the $T_9$ allele occurs only in the haplotypes that harbor the *OPN3490* G allele (Table 1). Of the 45 sires, 13 were heterozygous for *OPN3490*. Of these, seven had no evidence of any QTL segregating near *OPN*, one was significant for a QTL centromeric of *OPN* (M-IV-8), two (N-II-6 and N-III-3) were significant for QTL near 67 cM, and three (L-II-15, L-II-17, and L-II-4) appeared to segregate for two QTL in the region (Table 4).

**Additional Analyses Including *OPN* SNPs.** The *OPN* SNPs were integrated into the LK map, and LOKI was used to reanalyze the data (Fig. 1). Including these SNPs in the map did not affect the support or position of the FP QTL but increased support for the PP QTL by 103 units (BF, 332 vs. 229) and positioned the QTL directly over *OPN*. MY initially had no evidence of a QTL, but inclusion of the *OPN* SNPs produced a peak (BF, 6.0) directly over *OPN*. A third LOKI analysis, which included both *DGAT1* and *OPN3907* as fixed effects, completely eliminated the QTL support for both FP and MY. Support for the PP QTL at *OPN* was also eliminated, but support for the PP QTL near the casein cluster (88–90 cM) increased (BF, 248), indicating that this QTL is real. Fitting *OPN3907* as a fixed effect also revealed a PY QTL at 84 cM (BF, 83), within the same marker interval as the PP QTL (data not shown).

The *OPN* SNPs were integrated into the haplotypes, and the combined LK/LD analyses were repeated (Fig. 2). Large increases in lod scores were obtained between *OPN3492* and *OPN3907* for PP (from 19.6 without *OPN* SNPs to 23.2 with *OPN* SNPs) and MY (3.7 to 6.2) and between *OPN3490* and *OPN3492* for FP (5.5 to 11.4). These lod values were the maximum (for FP) or close to the maximum (for PP and MY) detected on the

**Table 1. Effect of *OPN3907* genotypes and *OPN* haplotypes on milk PTAs**

| | *OPN3907* | | | Haplotypes | | | | |
|---|---|---|---|---|---|---|---|---|
| | $T_9/T_{10}$ | $T_{10}/T_{10}$ | *P* | $GAGT_{10}$ | $TGAT_{10}$ | $GGGT_{10}$ | $GGGT_9$ | *P* |
| MY | 45.29 | −72.93 | 0.014 | −110.85 | −60.65 | −26.72 | −7.60 | 0.272 |
| FY | −5.93 | −1.95 | 0.031 | −1.92 | −1.72 | −2.77 | −7.34 | 0.062 |
| PY | −3.70 | −1.64 | 0.124 | −2.05 | −1.38 | −1.95 | −4.77 | 0.172 |
| FP | −0.0318 | 0.0036 | 1.36E-06 | 0.0097 | 0.0024 | −0.0066 | −0.0294 | 2.27E-04 |
| PP | −0.0221 | 0.0021 | 6.62E-14 | 0.0047 | 0.0017 | −0.0046 | −0.0202 | 3.27E-09 |

Haplotype order is *OPN3379-OPN3490-OPN3492-OPN3907*.

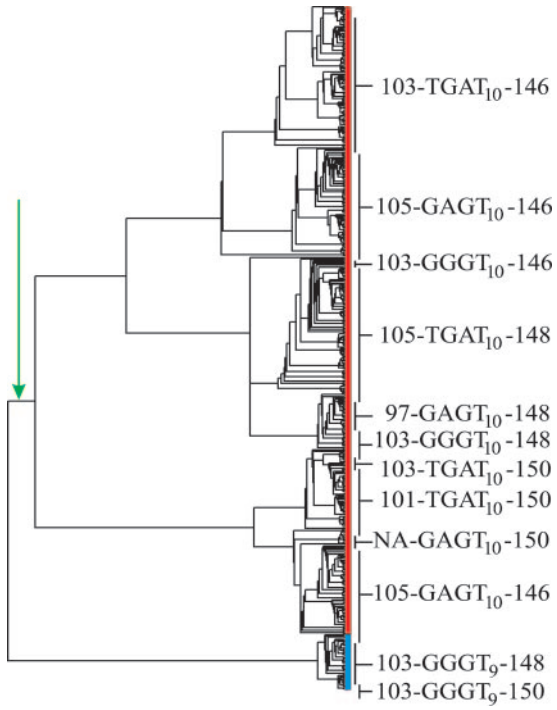


**Fig. 3.** Unweighted pair group method with arithmatic mean (UPGMA) haplotype dendrogram constructed at the *OPN3492–OPN3907* interval midpoint reveals two haplotype clusters at the first node (green arrow) uniquely defined by *OPN3907* alleles. Red (blue) bars indicate haplotypes containing $T_{10}$ ($T_9$). Black bars indicate clusters containing core *BMS2508-OPN3379-OPN3490-OPN3492-OPN3907-MNB175* haplotypes. The undefined *BMS2508* allele is identified as NA.

chromosome. However, lod scores for FY and PY in the vicinity of *OPN* did not increase (results not shown). Estimates of the proportion of variance in PTA (adjusted for *DGAT1*) explained by the QTL at the position of *OPN* for PP, FP, and MY were 44%, 19%, and 14%.

Fig. 3 shows the haplotype dendrogram with IBD estimated between *OPN3492* and *OPN3907*. Four main *OPN* haplotype families are represented within the phylogenetic tree: $TGAT_{10}$, $GAGT_{10}$, $GGGT_{10}$, and $GGGT_9$ (*OPN3379-OPN3490-OPN3492-OPN3907*), with frequencies of 53.3%, 19.7%, 15.7%, and 10.1%. The maximum lod of 23.2 for PP was obtained at the first node of this tree, which partitions haplotypes into two functionally homogeneous clusters based on alternate *OPN3907* alleles. Maximization of the likelihood within an interval flanked by *OPN3907* and with functionally distinct haplotypes defined exclusively by *OPN3907* alleles provides crucial evidence for the causality of *OPN3907* (22, 31).

Three additional analyses were performed to examine evidence for the causality of *OPN3907*. First, individual marker analyses indicated that only *OPN3907* was significant for PP (lod = 12.2; nominal $P$ = 6.6E-14) and FP (lod = 3.0; $P$ = 2.0E-4). Next, a combined LK/LD analysis for PP was performed in which *OPN3907* was fit as a fixed effect in the mixed QTL model. This resulted in complete elimination of support for any QTL in the vicinity of *OPN* (Fig. 2). Finally, assuming the QTL critical region to be 420-kb and the average SNP density in the bovine genome to be 0.85/kb, the probability there is no other mutation concordant with the eight sires' QTL genotypes is $P_0 \approx 0.40$. This is an upper bound, because the four segregating sires all shared the $T_9$ allele IBD. It would require independently sampling 10 heterozygotes and four homozygotes to increase $P_0$ to 0.95. These results, together with the segregation, ANOVA, and LOKI analyses for the *OPN* SNPs, strongly support the identity of *OPN3907* as the causal mutation underlying the PP QTL in interstitial BTA6.

```
BOVINE ACACCTTTTAAATGCCCATTAAAGTTTTTTTTTTTAATATAGA
CANINE ATACCTTTTAAATACCTAGTAAAGTTTTTTTTTTTAATATGGA
HUMAN  ACACCTTTTAAATACCTAGTAAAGTGTTTTTT..AATACAGA
CHIMP  ACACCTTTTAAATACCTAGTAAAGTGTTTTTT..AATACAGA
MOUSE  ATACCTCTTAACTATTCAGTAACGTTTT...GTTAATACAAA
```

**Fig. 4.** Multiple sequence alignment of the *OPN3907* region.

The entire bovine *OPN* sequence was aligned with the human and chimp genomic sequences by using ALIGN-M (32) (data not shown). The alignment was characterized by regions of moderate to high conservation with intervening stretches of no conservation, as expected from noncoding DNA. However, the motif containing *OPN3907* appears to have moderate conservation among bovine, dog, human, chimp, and mouse (Fig. 4). No transcription factor-binding sites known to be involved with mammary gland differentiation or development (www.cbil.upenn.edu/tess) were detected within the region.

## Discussion

We fine mapped a QTL affecting PP to a small interval on BTA6 in the vicinity of *BM143*. Examination of the genes in the region with conserved synteny on HSA4 identified *OPN* as an ideal functional candidate gene for this QTL. We identified four sires segregating for this QTL that were likely to be IBD for the detrimental QTL allele. Initial sequencing efforts focused on the promoter to polyadenylation signal region of *OPN*. Although we found three SNPs in this region, none were concordant with the segregation status of the sires. Higashibata *et al.* (33) demonstrated there were regulatory elements upstream of mouse *OPN* that controlled tissue-specific gene expression. They identified a region between −5,505 and −3,156 bp from the mouse *OPN* promoter that contained a mammary gland-specific regulatory element. We next obtained ≈5 kb of sequence upstream of the bovine *OPN* gene. Six additional SNPs were identified in this region, and all but *OPN3907* were excluded as causal based on concordance with the eight sire QTL genotypes. Our sequence alignment places the *OPN3907* ortholog at position −1,298 bp from the mouse *OPN* promoter, which does not appear to be within the mouse mammary-specific regulatory region (33).

Among sires born within the last 50 years, the MY-increasing $T_9$ allele has been decreasing in frequency, suggesting that either the allele is being selected against or the allele is being lost due to random drift. Because the U.S. Holstein population has been strongly selected for increased milk production over this period, we examined the effect of *OPN3907* on 20 body-conformation traits using PTAs generated by the Holstein Association. The only trait with a significant association was rear legs rear view ($P$ = 0.024), indicating that this mutation probably does not affect any of the conformation traits under selection. Further, *OPN3907* was not associated with somatic cell score, productive life, or daughter pregnancy rate PTAs (data not shown). Because the decrease in the $T_9$ allele frequency appears to be too rapid to be due to drift alone, we conclude the decrease is most likely due to selection on closely linked MY QTL that are in phase repulsion with the $T_9$ allele.

To test whether the QTL region on BTA6 has been under selection for milk production, LD was measured by using Hedrick D′ (34) between all pairs of markers for the son's maternal chromosomes ($n$ = 958). There was high LD between closely linked markers with the average D′ = 0.59 for markers ≤1 cM apart and ≥0.37 for markers ≤10 cM apart. The LD degraded with distance and reached an equilibrium of ≈0.14 at ≥50 cM between markers (Fig. 6, which is published as sup-

porting information on the PNAS web site). These results are similar to the LD estimates found between syntenic markers in a Dutch Holstein dairy cattle population (25). LD values were also compared between the 78 pairs of markers (excluding the *OPN* SNPs) within the 56- to 62-cM QTL region with LD values between 83 pairs of markers in the non-QTL harboring chromosomal regions. There was higher LD in this QTL region than in the non-QTL-harboring regions, with average D′s over distances of 0.5, 2, 3, 4, and 5 cM in the QTL (non-QTL) region of 0.55 (0.52), 0.51 (0.43), 0.47 (0.48), 0.61 (0.40), and 0.45 (0.42) (Fig. 6). The average D′s in the QTL and non-QTL regions were $0.54 \pm 0.02$ and $0.46 \pm 0.02$. A two-way ANOVA of D′ values indicated that LD differed between the QTL- and non-QTL-harboring regions ($P < 0.01$) and suggested an interaction between region and genetic distance ($P < 0.09$), indicating that the *OPN* region of BTA6 is under direct selection.

Although we have not proven that the *OPN3907* mutation is the causal mutation underlying the PP QTL in this region of BTA6, several lines of evidence point to causality:

(*i*) *OPN* is involved in the developmental regulation of the mammary gland in rodents, and regions upstream of the gene contain tissue-specific regulatory elements.

(*ii*) *OPN3907* is the only polymorphism within *OPN* that is concordant with the segregation status of the eight sequenced sires, and the probability that there is no other mutation concordant with the QTL genotypes of these sires within the 420-kb region is ≈0.40.

(*iii*) Alignment of the region harboring the *OPN3907* mutation among five mammalian species revealed a moderate conservation of the $T_n$ motif, suggesting that the mutation may lie within an *OPN* regulatory element.

(*iv*) Including the *OPN* SNPs in the marker map increased the statistical support for the MY, FP, and PP QTL directly over *OPN*.

(*v*) The joint LK/LD analysis positioned the PP QTL in the interval *OPN3492--OPN3907*, and the haplotype phylogeny produced two distinct clusters partitioned by *OPN3907* alleles.

(*vi*) Including *OPN3907* as a fixed effect completely eliminated evidence for this QTL.

The LD analyses produced distinct peaks at 62 cM (lod = 23.4, 8.5, and 5.2 for PP, FP, and MY) and at 65 cM (lod = 16.7, 4.3, and 6.5) within 9 cM of *OPN*. Reanalysis of the data after removing markers with <1,000 informative meioses revealed that QTL localizations were not artifacts of marker informativeness, and we conclude there are several milk-trait QTL located in interstitial BTA6. Consequently, the large effect of *OPN3907* (44% of the variance in PP PTA after accounting for *DGAT1*) is likely due to a confounding of closely linked PP QTLs in coupling phase and the strong LD within this region (Fig. 6). Of the sires segregating for at least one QTL, only L-I-1, L-II-14, L-II-15, and M-III-9 were heterozygous at *OPN3907*. However, M-III-9 was also detected to be segregating for the QTL at 65 cM with a minor peak at 57 cM, the location of *OPN*. The differences in estimated QTL location in the vicinity of *OPN* (Table 4) are likely due to the presence of several milk-trait QTL in various LK phases.

Our apparent positional cloning of the first milk-trait QTL on BTA6 should facilitate the discovery of the remaining QTL on this chromosome. Nevertheless, the close LK between QTL located both near *OPN* (57 cM) and the casein loci (88–90 cM) and the strong LD on this chromosome suggests that several populations must be examined to partition genetic variation into families that segregate for individual QTL. It remains to be demonstrated that *OPN3907* lies within a regulatory element. That the motif harboring *OPN3907* appears to be conserved among mammals but is not retrieved by a query of regulatory element databases suggests the element may be novel.

We thank T. J. Lawlor (Holstein Association U.S.A., Brattleboro, VT) for providing conformation trait data and the contributing artificial insemination organizations for donating semen to the Cooperative Dairy DNA Repository.

1. Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A., Sargeant, L., Sorensen, A., Steele, R., Zhao, X., *et al.* (1995) *Genetics* **139,** 907–920.
2. Khatkar, M., Thomson, P. C., Tamen, I. & Raadsma, H. W. (2004) *Genet. Sel. Evol.* **36,** 163–190.
3. Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., *et al.* (2002) *Genome Res.* **12,** 222–231.
4. Winter, A., Krämer, W., Werner, F.A.O., Kollers, S., Kata, S., Durstewitz, G., Buitkamp, J., Womack, J. E., Thaller, G. & Fries, R. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 9300–9305.
5. Bovenhuis, H. & Schrooten, C. (2002) *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production*, *Communication No. 09–07, August 19–23, 2002* (7WCGALP, Montpellier, France).
6. Ron, M., Kliger, D., Feldmesser, E., Seroussi, E., Ezra, E. & Weller, J. (2001) *Genetics* **159,** 727–735.
7. Freyer, G., Kühn, C., Weikard, R., Zhang, Q., Mayer, M. & Hoeschele, I. (2002) *J. Anim. Breed. Genet.* **119,** 69–82.
8. Olsen, H. G., Lien, S., Svendsen, M., Nilsen, H., Roseth, A., Aasland Opsal, M. & Meuwissen, T. H. E. (2004) *J. Dairy Sci.* **87,** 690–698.
9. Cohen, M., Reichenstein, M., Everts-van der Wind, A., Heon-Lee, J., Shani, M., Lewin, H. A., Weller, J. I., Ron, M. & Seroussi, E. (2004) *Genomics* **84,** 374–383.
10. Olsen, H. G., Lien, S., Gautier, M., Nilsen, H., Roseth, A., Berg, P. R., Sundsaasen, K. K., Svendsen, M. & Meuwissen, T. H. E. (2005) *Genetics* **169,** 275–283.
11. Denhardt, D. T. & Guo, X. (1993) *FASEB* **7,** 1475–1482.
12. Rittling, S. R. & Novick, K. E. (1997) *Cell Growth Diff.* **8,** 1061–1069.
13. Nemir, M., Bhattacharyya, D., Li, X., Singh, K., Mukherjee, A. B. & Mukherjee, B. B. (2000) *J. Biol. Chem.* **275,** 969–976.
14. D'Cruz, C. M., Moody, S. E., Master, S. R., Hartman, J. L., Keiper, E. A., Imielinski, M. B., Cox, J. D., Wang, J. Y., Seung, I. H A., Keister, B. A., *et al.* (2002) *Mol. Endocrinol.* **16,** 2034–2051.
15. Ashwell, M. & Van Tassell, C. (1999) *J. Dairy Sci.* **82,** Suppl. 1, 54.
16. Schnabel, R. D., Taylor, J. F. & Derr, J. N. (2003) *Cytogenet. Genome Res.* **102,** 59–64.
17. Green, P., Falls, K. & Crooks, S. (1990) CRIMAP Documentation (Washington Univ. School of Medicine, St. Louis, MO).
18. Thallman, R. M., Bennett, G. L., Keele, J. W. & Kappes, S. M. (2001) *J. Anim. Sci.* **79,** 26–33.
19. Thallman, R. M., Bennett, G. L., Keele, J. W. & Kappes, S. M. (2001) *J. Anim. Sci.* **79,** 34–44.
20. Seaton, G., Haley, C. S., Knott, S. A., Kearsey, M. & Visscher, P. M. (2002) *Bioinformatics* **18,** 339–340.
21. Heath, S. C. (1997) *Am. J. Hum. Genet.* **61,** 748–760.
22. Blott, S., Kim, J.-J., Moisio, S., Schmidt-Küntzel, A., Cornet, A., Berzi, P., Cambisano, N., Ford, C., Grisart, B., Johnson, D., *et al.* (2003) *Genetics* **163,** 253–266.
23. Churchill, G. A. & Doerge, R. W. (1994) *Genetics* **138,** 963–971.
24. Visscher, P. M., Thompson, R. & Haley, C. S. (1996) *Genetics* **143,** 1013–1020.
25. Farnir, F., Coppieters, W., Arranz, J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., *et al.* (2000) *Genome Res.* **10,** 220–227.
26. Meuwissen, T. H. & Goddard, M. E. (2001) *Genet. Sel. Evol.* **33,** 605–634.
27. Johnson, D. L. & Thompson, R. (1995) *J. Dairy Sci.* **78,** 449–456.
28. Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J.-J., Kvasz, A., Mni, M., Simon, P., Frere, J.-M., Coppieters, W., *et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101,** 2398–2403.
29. Ashwell, M., Van Tassell, C. & Sonstegard, T. S. (2001) *J. Dairy Sci.* **84,** 2535–2252.
30. Drenkard, E., Richter, B. G., Rozen, S., Stutius, L. M., Angell, N. A., Mindrinos, M., Cho, R. J., Oefner, P. J., Davis, R. W. & Ausubel, F. M. (2000) *Plant Physiol.* **124,** 1483–1492.
31. Kim, J.-J. & Georges, M. (2002) *Asian-Aust. J. Anim. Sci.* **15,** 1250–1256.
32. Van Walle, I., Lasters, I. & Wyns, L. (2004) *Bioinformatics* **20,** 1428–1435.
33. Higashibata, Y., Sakuma, T., Kawahata, H., Fujihara, S., Moriyama, K., Okada, A., Yasui, T., Kohri, K., Kitamura, Y. & Nomura, S. (2004) *J. Bone Miner. Res.* **19,** 78–88.
34. Hedrick, P. W. (1987) *Genetics* **117,** 331–341.