Multiple marker mapping of quantitative trait loci of Finnish dairy cattle by regression


Digital Object Identifier (DOI): 10.3168/jds.S0022-0302(97)75928-9

Link: Link to publication record in Edinburgh Research Explorer

Document Version: Publisher's PDF, also known as Version of record

Published In: Journal of Dairy Science

General rights Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Multiple Marker Mapping of Quantitative Trait Loci of Finnish Dairy Cattle by Regression

H. J. VILKKI, D.-J. de KONING, K. ELO, R. VELMALA, and A. MÄKI-TANILA
Agricultural Research Centre MTT, Institute of Animal Production, Section of Animal Breeding, FIN-31600 Jokioinen, Finland

ABSTRACT
A total of 453 bulls belonging to 11 half-sib families of Finnish Ayrshires were genotyped for six microsatellite markers on chromosome 9. The data were used in an attempt to map quantitative trait loci applying regression as a multimarker approach. For association analysis with a granddaughter design, the EBV for 12 traits were used: milk yield, protein yield, fat percentage, protein percentage, daughter weight, bull growth, calf mortality, days open, fertility treatments, nonreturn rate, SCC, and clinical mastitis. The empirical values of significance thresholds were determined using a permutation test on the experimental data. Although no significant effects were found, the results indicate some support for the existence of a locus on chromosome 9 that affects milk and protein yields.

(Key words: dairy cattle, interval mapping, quantitative trait, regression analysis)

Abbreviation key: QTL = quantitative trait locus.

INTRODUCTION
Almost all traits of dairy cattle that have economic interest are quantitative. That is, the observed phenotypes are continuously distributed and reflect the interaction of many quantitative trait loci (QTL) and environmental effects. Large-scale experiments using laboratory animals (12) and the existence of major loci in production traits, such as the double muscling gene in cattle or the Booroola high fertility gene in sheep, suggest that genes exist that have large or intermediate effects segregating for many traits in populations of farm animals. The segregation of such QTL could be detected by using analogous well-known probes for potential candidate genes [e.g., human halothane gene was used for pigs (13)] or by linked markers. When the number of known variable markers was still low, the choice was to quantify the association between the trait and a single marker using simple analysis of variance (15) or maximum likelihood (20). Paterson et al. (16) and Lander and Botstein (11) introduced interval mapping to determine the location of QTL by linkage analysis of maximum likelihood. This methodology has been successfully applied to different types of crosses of inbred lines. However, in outbred populations, these results might be biased if the markers (and thus intervals) for a linkage group were very different in their information content (6). Simultaneous use of multiple markers from a linkage group increases the sensitivity of the test statistic (6) and may eliminate bias in the estimated position and effect. Information from all (informative) markers for a linkage group has been utilized recently by Georges et al. (4) using maximum likelihood analysis in a half-sib population structure, but their method analyzes only one family at a time. A computationally less demanding alternative that uses a least squares approach with multiple markers has been suggested by Knott et al. (10). In addition, the regression method allows the analysis of several families with one analysis, and factors correcting for any systematic effects in the model could be included easily.

All of these methods suffer, however, from the difficulty in determining appropriate significance thresholds because of, for example, the large number of nonindependent tests used for QTL mapping. An empirical method used to estimate threshold values tailored to the experimental data has been recently presented by Churchill and Doerge (3), who suggested the use of a permutation test that involved repeated shuffling of the quantitative trait values and the generation of a random sample of the test statistic from an appropriate null distribution. Churchill and Doerge (3) presented examples of data obtained from maximum likelihood analysis and single marker analysis (t statistics), but regarded use of the permutation test to be feasible with any QTL mapping procedure.
Microsatellite loci have proved to be very useful genetic markers because of their high polymorphism and uniform distribution along the genome (8). Presently, two published linkage maps exist for cattle (1, 2) that cover most of the genome and contain a large selection of microsatellite markers. For cattle, microsatellites have been used in two previous attempts to map QTL using the granddaughter design. Ron et al. (17) used the single marker approach in a search for QTL affecting milk yield traits using 10 unlinked microsatellite loci and the progeny-testing results from 91 sons of 7 grandsires. Substitution effects ($P < 0.025$) were claimed for the paternal alleles at one microsatellite locus on milk yield and protein yield in one and two of the families, respectively. Georges et al. (4) used multiple marker mapping by maximum likelihood analysis in a study of Holstein dairy cattle with 14 grandsires and a total of 1518 sons. In that study, 159 microsatellite markers were typed and assigned to 29 synteny groups; five milk yield traits were analyzed using results of progeny testing. Using an LOD (logarithm of odds) score $>3$ as a significance threshold resulted in the assignment of QTL to the chromosomes 1, 6, 9, 10, and 20.

This study searched for QTL of Finnish Ayrshire dairy cattle by regression analysis using multiple markers. We used microsatellite loci on chromosome 9 and trait data obtained from the national animal (or sire) model evaluation in a granddaughter design. Chromosome 9 was chosen as an example because of the even distribution of published microsatellites and the previous indication of the existence (4) of a QTL affecting milk yield. To determine appropriate threshold values for this data, a permutation test was performed on each result of mapping.

**MATERIALS AND METHODS**

**Families**

Eleven extensively used AI bulls from dual purpose cattle of the Finnish Ayrshire breed were chosen to be the grandsires in a granddaughter design (Table 1). The oldest pedigrees, especially grandsires 1 and 2, included a selection bias caused by the unavailability of semen from the bulls that were ranked low after progeny testing. The EBV of traits were based on the recordings with the mean number of daughters per son varying from 147 to 1140.

The DNA was extracted from samples of frozen semen essentially as described by Zadworny and Kuhnlein (22). Quantitative trait scores were obtained from the national animal (or sire) model for cow (bull) evaluation of May 1995 for the following traits: milk yield, protein yield, fat percentage, protein percentage, daughter weight, bull growth, calf mortality, days open, fertility treatments, nonreturn rate, SCC, and clinical mastitis. For each trait, the BLUP for EBV of a bull was used. Use of daughter deviations would probably have been more correct, but, because all the bulls had well over 100 daughters (except for 5 bulls with 78 to 95 daughters), the use of EBV would not be affected by weighting in the least squares analysis.

**Genotyping**

The microsatellites (Table 2) to be typed were chosen to span the length of chromosome 9 in the published bovine linkage map (1) and Cattle Genome Database IRF11 (W. Barendse, 1994, personal communication). Primers for the microsatellite loci were synthesized with ABI 392 DNA synthesizer (Applied Biosystems Inc., Foster City, CA). One primer for each locus was labeled with fluorescein during synthesis (FluorePrime™; Pharmacia, Uppsala, Sweden).

The 25-$\mu$l polymerase chain reactions included 200 $\mu$M of each dNTP, DynaZyme™ buffer (Finzymes, Helsinki, Finland), 50 ng of template DNA, and 1 U of DynaZyme™II DNA polymerase (Finzymes). Multiplexing conditions and primer concentrations are given in Table 2. The reactions were carried out in a PTC-100 Programmable Thermal Cycler (MJ Research Inc., MA) with an initial denaturation step at 94°C for 5 min; followed by 26 cycles of 30 s at 94°C, 1 min at 55 or 58°C, and 35 s at 72°C; and a final extension step at 72°C for 8 min.

The amplified products were separated on 6% denaturing polyacrylamide gels (ReadyMix, Pharmacia) using the A.L.F.™ DNA Sequencer (Pharmacia). For size determination, an internal size standard (Sizer; Pharmacia) was included in each lane. The gels were analyzed using Fragment Manager 1.1 (Pharmacia).

Potential genotyping errors were screened for by reanalyzing all double recombinant genotypes.

**Statistical Analysis**

The bulls to be genotyped were chosen according to the granddaughter design (21) in which the sons of a grandsire are genotyped, and the association analysis is based on the means of the records from the daughters of each son. Although the size of the effect of a QTL is halved when measured among daughters, the
TABLE 1. The grandsires, number of studied sons by grandsire, number of granddaughters per grandsire, and number of informative sons per locus.

<table>
<thead>
<tr>
<th>Grandsire ID</th>
<th>Sons</th>
<th>Granddaughters</th>
<th>CSSM 56 (no.)</th>
<th>TGLA 73 (no.)</th>
<th>UWCA 9 (no.)</th>
<th>CSSM 25 (no.)</th>
<th>ETH 225 (no.)</th>
<th>INRA 136 (no.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33090</td>
<td>30</td>
<td>33,452</td>
<td>17</td>
<td>25</td>
<td>19</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>33787</td>
<td>29</td>
<td>30,794</td>
<td>8</td>
<td>22</td>
<td>25</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>34740</td>
<td>58</td>
<td>19,960</td>
<td>...</td>
<td>48</td>
<td>39</td>
<td>...</td>
<td>43</td>
</tr>
<tr>
<td>4</td>
<td>34798</td>
<td>41</td>
<td>8,133</td>
<td>25</td>
<td>28</td>
<td>28</td>
<td>33</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>34872</td>
<td>50</td>
<td>11,519</td>
<td>26</td>
<td>46</td>
<td>38</td>
<td>22</td>
<td>...</td>
</tr>
<tr>
<td>6</td>
<td>35076</td>
<td>21</td>
<td>5,742</td>
<td>...</td>
<td>19</td>
<td>...</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>35142</td>
<td>82</td>
<td>17,079</td>
<td>...</td>
<td>59</td>
<td>55</td>
<td>57</td>
<td>...</td>
</tr>
<tr>
<td>8</td>
<td>35144</td>
<td>29</td>
<td>5,844</td>
<td>...</td>
<td>...</td>
<td>26</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>9</td>
<td>36022</td>
<td>29</td>
<td>5,091</td>
<td>14</td>
<td>20</td>
<td>22</td>
<td>15</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>36378</td>
<td>44</td>
<td>5,317</td>
<td>...</td>
<td>...</td>
<td>29</td>
<td>17</td>
<td>31</td>
</tr>
<tr>
<td>11</td>
<td>36386</td>
<td>40</td>
<td>6,292</td>
<td>23</td>
<td>...</td>
<td>...</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>453</td>
<td>149,223</td>
<td>113</td>
<td>267</td>
<td>255</td>
<td>226</td>
<td>221</td>
</tr>
</tbody>
</table>

1Identification.

The principle of interval mapping is to fit a model with the QTL effect at different positions between two adjacent markers. Haley and Knott (5) and Martinez and Curnow (14) have described a method that applies regression to estimate the location and the effect of a QTL using pairs of adjacent (informative) markers.

The multiple marker approach for half-sib families (10) with the granddaughter design was used with the following steps. First, the most likely linkage phases of the gametes of grandsire were determined. After screening which markers were informative for each of the sons of the grandsire, the probability of the son inheriting the first grandsire gamete was calculated for positions at 1-cM intervals throughout the chromosome. This probability was computed using the estimated recombination fraction between the two closest flanking informative markers. For instance, assume a grandsire that is heterozygous for two markers with the chromosome phases $AB/ab$. The recombination fraction between the markers is $R$. If $r$ is the distance between the marker $A$ and the putative QTL and if the grandsire chromosome $AB$ has the QTL allele with a larger effect, then the probability can be computed that a son having inherited the $AB$ combination would also have the QTL allele with the larger effect. Assuming no interference among loci, the probability is $1 - r(R - r)/(1 - R)(1 - 2r)$. The complementary event is the existence of a large QTL allele with the marker combination $ab$.

TABLE 2. The microsatellite loci used with the number of alleles observed, polymorphism information content (PIC), the percentage of informative sons in heterozygous families and the conditions of polymerase chain reactions for each locus.1

<table>
<thead>
<tr>
<th>Locus</th>
<th>Alleles (no.)</th>
<th>PIC</th>
<th>Informative sons (%)</th>
<th>Annealing temperature (°C)</th>
<th>Multiplex Concentration (µM)</th>
<th>Primer Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSSM 56</td>
<td>5</td>
<td>0.594</td>
<td>51.6</td>
<td>55</td>
<td>TGLA 73</td>
<td>0.5</td>
</tr>
<tr>
<td>TGLA 73</td>
<td>6</td>
<td>0.675</td>
<td>78.5</td>
<td>55</td>
<td>CSSM 56</td>
<td>0.5</td>
</tr>
<tr>
<td>UWCA 9</td>
<td>7</td>
<td>0.729</td>
<td>70.2</td>
<td>55</td>
<td>INRA 136</td>
<td>1.0</td>
</tr>
<tr>
<td>CSSM 25</td>
<td>3</td>
<td>0.431</td>
<td>57.2</td>
<td>55</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ETH 225</td>
<td>7</td>
<td>0.787</td>
<td>75.7</td>
<td>55</td>
<td>...</td>
<td>1.0</td>
</tr>
<tr>
<td>INRA 136</td>
<td>4</td>
<td>0.615</td>
<td>73.3</td>
<td>58</td>
<td>UWCA 9</td>
<td>0.3</td>
</tr>
</tbody>
</table>

1Source of primer sequences was Barendse et al. (1) for CSSM 56, TGLA 73, UWCA 9, CSSM 25, and ETH 225 and Vaiman et al. (19) for INRA 136.

2In the published sequence of INRA 136, the last base of the reverse primer should be a G instead of a C.
Similarly, for the marker pair \( Ab \), the probability is 
\[
(1 - r)(R - r)/R(1 - 2r),
\]
the complementary event of which is the QTL associated with the pair \( aB \). For a locus outside the marker series, the probability of inheriting the QTL allele with the larger effect is the assumed recombination fraction between the QTL and the closest informative marker locus, as suggested by Knott et al. (10). Because transmission probabilities are calculated for each position on the chromosome, the degrees of freedom are equal at each point. This approach makes possible a full use of the data when families with poorer marker information can be included in the analysis for these regions. However, because the analysis outside the segment bracketed by informative markers is based on single marker associations rather than interval mapping, some caution should be used when interpreting the results (regarding position) toward the ends of the chromosome. The test statistic is constant over the regions outside the markers because the analysis for these regions is based purely on single markers, and the effect and location are confounded. To see how this situation affects the results, we also analyzed our data including only informative intervals for each individual (pure interval mapping).

Each position was assessed for the existence of a QTL by regressing the trait score on probability; the analysis gave an estimate of both the location and effect of the QTL. The regression coefficients were estimates of the substitution effect at the position considered. The model assumed one QTL with an additive effect and no interaction because half-sib analysis could not detect dominance and because epistasis would require the allowance for several QTL. Conversely, one of the appealing properties of the regression approach is that it could quite easily be extended to cover multiple QTL (9, 23).

An intraclss regression model, which was similar to that of Knott et al. (10), was used:

\[
Y_{ijk} = \mu + a_i + b_iX_{ijk} + e_{ijk}
\]

where

\[
Y_{ijk} = \text{EBV of a bull } k, \text{ son of grandsire } i, \text{ marker genotype } j;
\]
\[
\mu = \text{overall mean};
\]
\[
a_i = \text{effect because of grandsire } i;
\]
\[
b_i = \text{regression coefficient within grandsire } i;
\]
\[
X_{ijk} = \text{probability of the large QTL allele being transmitted from the grandsire } i \text{ given the pair of detected flanking markers } j \text{ of the son } k; \text{ and}
\]
\[
e_{ijk} = \text{residual effect}.
\]

The QTL alleles of different families may be different in their effect, and, more likely, their linkage phase with markers may differ between families. With an overall analysis of the families, the significance of QTL results can be inferred from the variation because of regression within families. The pooled mean square that was due to regression within grandsires was computed, and its ratio to the residual mean square \((F \text{ ratio})\) was used as a test statistic. The peak value of this statistic appears at the most likely position of a QTL on the chromosome.

The methodology for the determination of the significance thresholds of a test statistic is still under debate. We used a permutation approach, as suggested by Churchill and Doerge (3). The trait scores (for the permutation test, the EBV were corrected for the EBV of sires) were shuffled, and genotypes were retained. The \(F \) ratio was calculated at each analysis point at 1-cM intervals, and this procedure was repeated 10,000 times. Permuted data were sorted to obtain 0.1, 1, and 5% nominal significance thresholds \((P_n)\). Overall significance thresholds \((P_o)\) were calculated assuming eight independent tests, \(P_o = 1 - (1 - P_n)^8\). Data on the 12 separately recorded traits were utilized. However, some traits are genetically highly correlated (>0.6), and the actual number of independent traits was thought to be 8: i.e., yield (milk yield and protein yield), milk content (fat percentage and protein percentage), daughter weight, bull growth, calf mortality, cow fertility (days open and fertility treatments), nonreturn rate, and udder

### TABLE 3. Map distances (Haldane) between microsatellite loci on chromosome 9, calculated from data on male meioses in the Finnish Ayrshire.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Informative meioses</th>
<th>Distance</th>
<th>Reference value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSSM 56</td>
<td>75</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>TGLA 73</td>
<td>182</td>
<td>33</td>
<td>25</td>
</tr>
<tr>
<td>UWCA 9</td>
<td>128</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>CSSM 25</td>
<td>107</td>
<td>48</td>
<td>23</td>
</tr>
<tr>
<td>ETH 225</td>
<td>135</td>
<td>16</td>
<td>. .</td>
</tr>
</tbody>
</table>

1The number of coinformative meioses between each adjacent marker pair.

2The respective reference values (Kosambi) are from Barendse et al. (1).
RESULTS AND DISCUSSION

The microsatellites used proved to be moderately informative (Table 2). Although the marker loci had 3 to 7 alleles, the percentage of informative sons from heterozygous sires (sons with a genotype different from the sire genotype) ranged from only 52 to 79%. Three grandsires were heterozygous for all six markers, but 3 were heterozygous for only three markers (Table 1). Consequently, at each analysis point, only approximately half of the progeny was informative.

The distances between loci in our data (Table 3) were somewhat different from those shown by the map of Barendse et al. (1). This result could be because of the Haldane mapping function or because our data were based on only male meioses in a single breed.

The threshold values for different risk levels obtained from the permutation tests were almost identical, regardless of the trait. The $F$ values at experimentwise thresholds varied from 2.71 to 2.82 (1% nominal level/10,000 permutations) or from 3.21 to 3.48 (0.1% level). We calculated the 0.1% nominal level, because, assuming eight independent traits were analyzed. The nominal 5% experimentwise threshold equals the 1% comparisonwise threshold (comparisonwise 5% threshold is shown as a dashed line). Marker locations are shown in centimorgans above the X-axis.

The result suggests that, in most cases, using a 1% nominal level, empirical significance with 10,000 rounds of permutation for one trait is sufficient, which would shorten the computation time considerably. Nevertheless, routine use of a permutation test with 10,000 shuffled analyses in connection with the multimarker regression is feasible; for example, 48 min were needed to analyze 10,000 permutations for an average family of 40 sons and 190 min for the whole data file (453 bulls). We used an IBM-compatible computer (486DX2 and 66 MHz) under DOS. The test statistic was calculated for each permutation at 1-cM intervals using the precision of six decimal digits. The permutations were performed on a PC, and results were sorted on a UNIX workstation (IBM RISC System/6000). The sorting of 10,000 permutations took 15 to 30 min and 4 to 12 h time for experimentwise and comparisonwise tests, respectively (the exact usage of time depends on the sorting algorithm used). As a comparison, Churchill and Doerge (3) used 23 h to generate 1000 maximum likelihood analyses of permuted data (12 marker loci and 333 individuals) on a SPARC IPX Workstation.

To utilize the information for marker-assisted selection, for example, analysis must also be performed at the family level. As a rough guideline to the significance of $F$ values from the within-family analysis, Scheffé's (18) result was used: the effect is significant if it exceeds $((\text{n.o. grandsires} - 1)F)^{0.5}$ for the respective risk level, where $F$ is the threshold value for the across-family analysis at the respective risk level. Therefore, we concluded that a QTL effect in an individual family is significant with overall $P <0.01$, if the $F >5.7$ (experimentwise) or $>5.3$ (comparisonwise).

Figure 1 shows the test statistics of full regression analysis for multiple markers for milk and protein yields. The results from a pure interval analysis, omitting regions outside informative intervals, were
essentially similar, except that both the test statistic and significance levels were higher (about 0.5 units). Based on the permutation test (10,000 permutations) and taking into account the repeated tests, no effect across families was found for any of the 8 traits. However, some support was found for the presence of a QTL affecting milk yield. The $F$ ratio reached its highest value, 2.196, at 64 cM. At that position, the comparisonwise threshold level of 1% ($P = 0.05$) was 2.25. At 65 cM is marker UWCA9, which in the map of Georges et al. (4) would reside in the same region where the QTL for milk and protein yields were found in their (4) study. Moreover, test statistics were $F > 5.3$ for three families, two of which (33090 and 36386) were related. Grand sire 33090 was the sire of grandsires 36378 and 36386. According to their phases, the chromosomes inherited from 33090 were different at all markers; if the sire genotype is marked ABCDEF/abcdef, 36378 inherited a chromosome abCDEF and 36386 a chromosome ABcdef. Thus, the two sons inherited different alleles for the putative QTL.

One point of discussion is the relatively small size of our families. The number of progeny needed to detect a QTL depends on QTL effect, marker spacing, distance between QTL and the closest marker, informative level of markers, and heritability of the trait. Georges et al. (4) considered single family analyses simulating a QTL with an effect of 1 standard deviation for 1000 pedigrees. For 50 sons per grandsire, effects were significant only for 40% of the cases; for 100 and 200 sons per grandsire, the scores were 90 and 100%, respectively. A situation more closely resembling our analysis (significance over several families) was simulated by Weller et al. (21). For 10 grandsires with 40 sons each (100 daughters per son), the statistical power varied between 2 and 94%, depending on the QTL effect and the heritability of the trait. For example, for a trait with 0.2 heritability and effect size of 0.3 standard deviation units, the statistical power was calculated (21) to be 78%. Seemingly, the sample size in our study may have been too small to obtain enough statistical power to detect QTL of reasonable magnitude, especially because only about half of the sons were informative at each analysis point.

We did not test the possibility of two QTL affecting the same trait, as has been suggested by Haley and Knott (5) and Martinez and Curnow (14). Haley et al. (7) suggested that their regression methods could be followed by computationally more demanding procedures, such as maximum likelihood, for the traits and areas where possible QTL appear, which could result in better estimates of the position of the QTL and the magnitude of its effect.

CONCLUSIONS

We have shown that interval mapping can be carried out using regression and multiple markers in half-sib pedigrees and that empirical significance threshold values can be quite easily found in this context by permuting the experimental data. With these tools, we were able to find some support for the presence on chromosome 9 of a QTL controlling milk and protein yields within an active breeding population.

ACKNOWLEDGMENTS

This research was partly supported by the Ministry of Agriculture and Forestry in Finland. We are grateful to Anneli Virta for the A.L.F. analyses, the Finnish AI associations for providing the sperm samples, and the Agricultural Computing Center of Finland for the progeny-testing results.

REFERENCES
