RESEARCH ARTICLE

DNA marker mining of ILSTS035 microsatellite locus on chromosome 6 of Hanwoo cattle

JUNG-SOU YEO[1,2], JEA-YOUNG LEE[3]* and JAE-WOO KIM[2]

[1]Department of Animal Science, [2]Institute of Biotechnology, and [3]Department of Statistics, Yeungnam University, Gyeongsan, Gyeongbuk 712-749, Korea

Abstract

We describe tests for detecting and locating quantitative trait loci (QTL) for traits in Hanwoo cattle. From results of a permutation test to detect QTL for marbling, we selected the microsatellite locus ILSTS035 on chromosome 6 for further analysis. *K*-means clustering analysis applied to five traits and nine DNA markers in ILSTS035 resulted in three cluster groups. Finally we employed the bootstrap test method to calculate confidence intervals using the resampling method to find major DNA markers. We conclude that the major markers of ILSTS035 locus on chromosome 6 of Hanwoo cattle are markers 235 bp and 266 bp.

[Yeo J.-S., Lee J.-Y. and Kim J.-W. 2004 DNA marker mining of ILSTS035 microsatellite locus on chromosome 6 of Hanwoo cattle. *J. Genet*. **83**, 245–250]

Introduction

Problems in detecting and locating quantitative trait loci (QTL) have received considerable attention over the past several years. A variety of methods have been developed to analyse quantitative-trait data (Weller 1986, Lander and Botstein 1989, Churchill and Doerge 1994). Many research groups (Hirano *et al.* 1998; Kim *et al.* 2000, 2003a,b) have intensively analysed linkage between markers and traits to identify chromosomal regions responsible for economically important traits such as meat quality and carcass length. Some traits such as 'double muscle' in cattle and RN in swine were revealed to be due to particular genes (McPherron and Lee 1997). Such identification of genes responsible for traits requires a huge amount of research, time, and some luck. If gene arrangement along chromosomes is determined completely or nearly completely, one can select gene candidates for traits very efficiently and speed up identification of the genes responsible for the traits. A common problem in all of these methods is the difficulty of determining appropriate significance thresholds (critical values) against which to compare test statistics (usually LOD scores or likelihood ratios) for the purpose of detecting QTL. Knott and Haley (1992) used simulation study for the distributional properties of likelihood ratio tests for QTL detection. They suggested that the chi-square approximation to the distribution of likelihood ratio test statistic is not reliable in many cases and requires further theoretical work. Churchill and Doerge (1994) proposed permutation tests to detect QTL in the genome. An introduction to the theory of permutation testing is provided by Good (1994).

In the work reported here, we tried a method based on the concept of permutation test (Good 1994), because major LOD scores do not have theoretical significance levels (critical value or *P* value). Ten thousand repetitions of the permutation process were used for critical value. A microsatellite locus, ILSTS035, was selected by permutation testing. This locus includes nine 'genes': DNA markers 210 bp, 215 bp, 230 bp, 235 bp, 240 bp, 245 bp, 255 bp, 260 bp and 266 bp. Next, the relations between the DNA markers and the economic trait were identified by *K*-means clustering analysis. Finally, we applied the bootstrap test (Efron 1987; Visscher *et al*. 1996) to calculate confidence intervals of QTL locations for traits. The

*For correspondence. E-mail: jlee@yu.ac.kr.

Keywords. LOD score; QTL; permutation test; *K*-means clustering; bootstrap method.

number of bootstrap samples for each DNA was 1000 and 95% confidence intervals were calculated for economically important traits.

Materials and methods

Animals and traits

One hundred and thirtyseven steers from 10 paternal half-sib families were used for linkage mapping and QTL from Hanwoo Improvement Center, National Agricultural Cooperation Federation, Korea. Daily weight gain was measured from birth to 720 days of age, and marbling scores were measured at slaughter, at 720 days of age. Marbling was scored as 19 degrees and classified by 1+, 1, 2 and 3 for market systems. The grading of the marbling scores, backfat thickness and the m. longissimus dorsi area were measured according to standards of the Korean Animal Products Grading Service.

Permutation tests

A permutation test in the simplest case is used to detect a location shift in data that are divided into two sets of observations. LOD scores (exceeding 3) for detecting and locating quantitative trait loci (QTL) from the Hanwoo marbling scores were selected, and are shown in table 1. However, LOD scores at which significance is declared cannot be obtained theoretically, therefore we applied the genomewise (experimentwise) permutation test (Churchill and Doerge 1994). We followed the five-step procedure (Good 1994, p. 20) for a permutation test: Step 1: Analyse the problem (hypothesis, distribution drawn, etc.). Step 2: Choose the test statistic (sum of observations in the first sample) which will distinguish the hypotheses. Step 3: Compute the test statistic for the original labelling of the observations. Step 4: Rearrange (permute) the labels and recompute the test statistic for the new labels. Repeat until you obtain the distribution of the test statistic for possible permutations. Step 5: Calculate the labels of significance using this permutation distribution of the statistic.

An empirical 100(1–*P*) percentile obtained by 10,000 repetitions of the permutation process was referred to as an estimated critical value of the genomewise significance level of *P*. The critical value of $P = 0.01$ was used to detect the presence of a QTL somewhere in the genome, so that the type I error rate may be 0.01 or less (table 1). Six loci, including ILSTS035, were selected.

After the permutation test, we needed to identify the major DNA markers in ILSTS035 based on economically important traits such as meat quality and carcass length.

K-means clustering methods

Grouping or clustering can provide an informal means of assessing dimensionality, identifying outliers, and suggesting interesting hypotheses concerning relationships. The *K*-means method, which was suggested by MacQueen (1967), is a nonhierarchical clustering technique. The process is composed of the following three steps: Step 1: Partition the items into *K* initial clusters. Step 2: Proceed through the list of items, assigning an item to the cluster whose centroid (mean) m_t ($t = 1, \ldots, k$) is nearest. Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item. Step 3: Repeat step 2 until no more reassignments take place.

Distance is usually computed using Euclidean distance with either standardized or unstandardized observation vectors X_i ($i = 1, \ldots, n$). That is, from ($p \times n$) data matrix X and variance–covariance matrix S:

$$X = \begin{bmatrix} x_{11}\, x_{12} \ldots x_{1n} \\ x_{21}\, x_{22} \ldots x_{2n} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ x_{p1}\, x_{p2} \ldots x_{pn} \end{bmatrix} = [X_1\, X_2 \ldots X_n]$$

$$S = \begin{bmatrix} s_1^2\ s_{12} \ldots s_{1p} \\ s_{21}\, s_2^2\ \ldots s_{2p} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ s_{p1}\ s_{p2} \ldots s_p^2 \end{bmatrix}.$$

The Euclidean distance between *i*th observation *j*th variable x_{ij} and *t*th clustering centroid mean m_t ($t = 1,2,\ldots, k$) depends on

$$d(X_i\, m_t) = [\sum_{j=1}^{p} (x_{ij} - m_{tj})^2]^{1/2},$$

and standardized Euclidean distance depends on

$$d(X_i\, m_t) = [\sum_{j=1}^{p} (\frac{x_{ij} - m_{tj}}{s_j})^2]^{1/2}.$$

Rather than starting by partitioning all items into *K* preliminary groups in step 1, we could specify *K* initial centroids and then proceed to step 2.

The results are shown in tables 2 through 4 with figure 2.

Bootstrapping BCa (bias-corrected and accelerated) analysis

Sampling with replacement of *n* individual observations created bootstrap samples. An observation consists of a marker genotype and a phenotype, so at each bootstrap sample, we drew, with replacement, *n* observations out of the pool of (*n*) original observations. Some records can appear more than once in a bootstrap sample, while others are not included at all. After determining the *n* bootstrap samples, the empirical central 90% and 95% confidence intervals of the QTL positions were determined by ordering the *n* estimates and taking the bottom and top 5th and 2.5th percentile, respectively. The bootstrap idea

is simply to replace the unknown population distribution with the known empirical distribution function. The bootstrap distribution for $\hat{q}-q$ is the distribution determined by generating $\hat{q}$ values which are determined by sampling independently with the replacement form empirical distribution, F_n. The bootstrap estimate of the standard error of $\hat{q}$ then becomes the standard deviation of the bootstrap distribution for $\hat{q}-q$.

It should be noted here that almost any parameter of the bootstrap distribution may serve as a 'bootstrap' estimate of the corresponding population parameter. We could consider the skewness, the kurtosis, the median, or the 95th percentile of the bootstrap distribution for $\hat{q}$. The basic idea behind the bootstrap is that the variability of q^* around $\hat{q}$ will be similar to the variability of $\hat{q}$ around s. There is good reason to believe this will be true for large sample sizes, since we see that as n grows larger, F_n becomes comparable to random sampling from F.

We have the following steps to produce BCa (bias-corrected and accelerated) bootstrap intervals: Step 1: Generate a sample of size n with replacement from the empirical distribution. Step 2: Compute q^*, the value of $\hat{q}$ obtained by using the bootstrap sample in place of the original sample. Step 3: Repeat steps 1 and 2 k times. By replicating steps 1 and 2 k times, we obtain a Monte Carlo approximation to the distribution of q^*. Let $\hat{q}^*_{(a)}$ indicate the $100 \times a$ th percentile of $B = 1000$ bootstrap replications $\hat{q}^*_{(1)}, \hat{q}^*_{(2)}, \ldots, \hat{q}^*_{(B=1000)}$. Step 4: The BCa interval endpoints are also given by percentiles of the bootstrap distribution. The percentiles used, however, depend on two numbers, $\hat{a}$ (acceleration) and Z_0 (bias correction).

The BCa interval of intended coverage $1–2a$ is given by

$$\text{BCa}; (\hat{q}_{\text{lo}}, \hat{q}_{\text{up}}) = \hat{q}^{*(a1)}, \hat{q}^{*(a2)},$$

where $a_1 = \Phi(\hat{Z}_0 + (\hat{Z}_0 + Z^{(a)})/[1-\hat{a}(\hat{x}_0 + Z^{(a)})])$, $a_2 = \Phi[\hat{Z}_0 + (\hat{Z}_0 + Z^{(1-a)})/(1-\hat{a}(\hat{x}_0 + Z^{(1-a)}))]$. Here $\Phi(\cdot)$ is the standard normal cumulative distribution function, $Z^{(a)}$ is the 100th percentile point of standard normal distribution. If $\hat{a}$ and $\hat{Z}_0$ equal zero, then the BCa interval is the same as the percentile interval; If $\hat{a}$ and $\hat{Z}_0$ are not equal to zero, then the BCa interval endpoints change. Bias correction $\hat{Z}_0$ is obtained from

$$Z_0 = \Phi^{-1}\left[\frac{\sum_{b=1}^{n} I(\hat{q}^*(b) < \hat{q})}{B}\right],$$

where Φ^{-1} is the inverse function of the standard normal cumulative distribution function.

Results and discussion

QTL methodology

LOD scores and the permutation test for detecting and locating quantitative trait loci (QTL) from the Hanwoo marbling scores are given in table 1. We selected several loci that had maximum LOD scores exceeding 3, which is generally considered significant (Chotai 1984). However, LOD scores at which significance is declared cannot be obtained theoretically; therefore we applied the genomewise (experimentwise) permutation test (Churchill and Doerge 1994). An empirical 100(1–P) percentile obtained by 10,000 repetitions of permutation process for each locus was referred to as an estimated critical value of the genomewise significance level of P. The critical value of $P = 0.01$ was used to detect the presence of a QTL somewhere in the genome so that the type I error rate may be 0.01 or less (table 1).

In table 1, AFR227 is not significant statistically, but other loci show very significant levels of P. In particular, ILSTS035 and BM4311 were demonstrated to be the best. The present work was an attempt at DNA marker mining of the ILSTS035 microsatellite locus on Hanwoo chromosome 6.

K-means clustering

One hundred and thirtyseven steers were used for the analysis. We analysed the ILSTS035 microsatellite locus on chromosome 6. Nine DNA markers were obtained (210 bp, 215 bp, 230 bp, 235 bp, 240 bp, 245 bp, 255 bp, 260 bp, 266 bp) as well as data on five economic traits, namely marbling score, daily gain, backfat thickness, m.

Table 1. Permutation test results based on marbling.

Trait	Locus	LOD score	*P* value*	Ratio of QTL variation (%)
Marbling score	BM3026	3.501999070	< 0.01	9.37
	BMS690	4.602844364	< 0.01	12.53
	ILSTS035	4.991694330	< 0.01	16.54
	BM4311	6.594214985	< 0.01	16.38
	BMS511	4.079206916	< 0.01	11.03
	AFR227	3.150859863	0.07235	19.47
	BMC4203	2.915809464	< 0.01	7.88

*Test statistic for the significance level is the sum of observations in the first sample (Good 1994).

longissimus dorsi area and carcass weight.

The *K*-means clustering analysis method applied to the five traits and nine DNA markers resulted in three cluster groups (table 2 and figure 1). From table 2 we can conclude that cluster 1 is a useful group for backfat thickness (high value = 0.33065), cluster 2 is a useful group for marbling score (high value = 1.341518), and cluster 3 is a useful group for carcass weight, daily gain and m. longissimus dorsi area. Figure 1 shows that cluster 1 has a great proportion of DNA markers 210, 230 and 245 bp, cluster 2 has a great proportion of markers 235 and 240 bp, and cluster 3 has a great proportion of markers 260 and 266 bp. Marker 240 bp however was seen in very few individuals (n = 4) and this number may not be sufficient for drawing conclusions.

Similarly, we recorded standardized mean results for the five economic traits compared with the DNA markers (table 3). Marker 235 bp presents a higher marbling score, markers 240 and 255 bp present higher daily gain, markers 210 and 235 bp present higher backfat thickness, marker 266 bp presents higher m. longissimus dorsi area, and markers 240 and 266 bp present higher carcass weight.

A summary of the results is given in table 4. Marker 210 bp is a useful one for backfat thickness; marker 235 bp for marbling score; and marker 266 bp for daily gain, m. longissimus dorsi area and carcass weight. Although marker 210 bp is important for backfat and marker 235 bp for marbling, the numbers of individuals are only 10 and 15, which may be insufficient for the conclusion. Therefore we decided to try bootstrap testing.

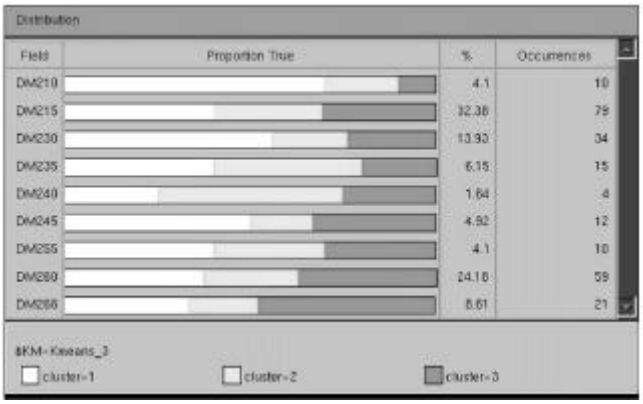

Figure 1. Clustering proportional analysis for DNA markers.

Bootstrap (BCa method) analysis

We applied the bootstrap testing method (Visscher *et al*. 1996) to calculate confidence intervals for finding major DNA markers. Bootstrap samples were created by sampling with replacement each individual DNA marker and trait. The number of bootstrap samples for each marker was 1000, and 95% confidence intervals of bootstrap testing were calculated for the five traits, i.e. marbling score, daily gain, backfat thickness, m. longissimus dorsi area and carcass weight (figures 2 through 6).

Figure 2 shows that marker 235 bp gives better interval (7.726 ~ 12.0667) and mean (9.8667) for marbling than others. In figure 3, we don't have an especially good confidence interval for daily gain. Figure 4 shows that marker 210 bp gives a lower confidence interval (3.8 ~ 5.5) for backfat thickness, which is good. In figures 5 and 6, we

Table 2. *K*-means clustering analysis.

Trait	Cluster 1 (58)*	Cluster 2 (36)	Cluster 3 (43)
Marbling score	– 0.673177	**1.341518**	– 0.215128
Daily gain	– 0.711536	0.08525	**0.888374**
Backfat thickness	**0.333065**	– 0.11327	– 0.354413
M. longissimus dorsi area	– 0.564381	0.169661	**0.619216**
Carcass weight	– 0.791399	0.172412	**0.923122**

*Numbers in parenthesis are numbers of individuals.

Table 3. Standardized mean results for the five traits and DNA markers of ILSTS035.

	DNA marker								
Trait	210 bp (10)*	215 bp (79)	230 bp (34)	235 bp (15)	240 bp (4)	245 bp (12)	255 bp (10)	260 bp (59)	266 bp (21)
Marbling score	– 0.4286	0.0804	– 0.1459	**0.5838**	0.1355	– 0.1981	– 0.134	0.0147	– 0.0356
Daily gain	– 0.5216	0.0741	– 0.2309	– 0.1715	**0.2109**	– 0.364	**0.2196**	0.0384	0.182
Backfat thickness	**0.8808**	0.1168	0.0597	**0.2417**	0.0689	– 0.407	– 0.0365	– 0.2142	0.0967
M. longissimus dorsi area	0.0483	0.0668	– 0.1997	0.1683	0.1968	– 0.5815	– 0.196	0.1646	**0.4375**
Carcass weight	– 0.5999	0.0594	– 0.2924	– 0.0201	**0.3722**	– 0.4761	– 0.1376	0.1091	**0.1997**

*Numbers in parenthesis are numbers of individuals.

Table 4. Clustering comparison between means and *K*-means mining results.

Cluster group	Mean result	*K*-means mining
Backfat thickness (cluster 1)	210 bp, 235 bp	210 bp, 230 bp, 245 bp
Marbling score (cluster 2)	235 bp	235 bp, 240 bp
Daily gain, M. longissimus dorsi area and carcass weight (cluster 3)	240 bp, 255 bp, 266 bp	260 bp, 266 bp

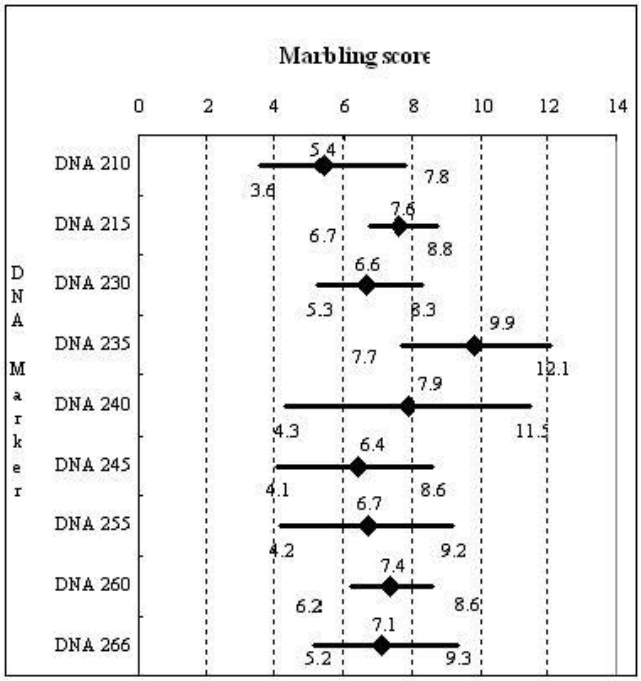

Figure 2. Bootstrap confidence intervals for marbling score.

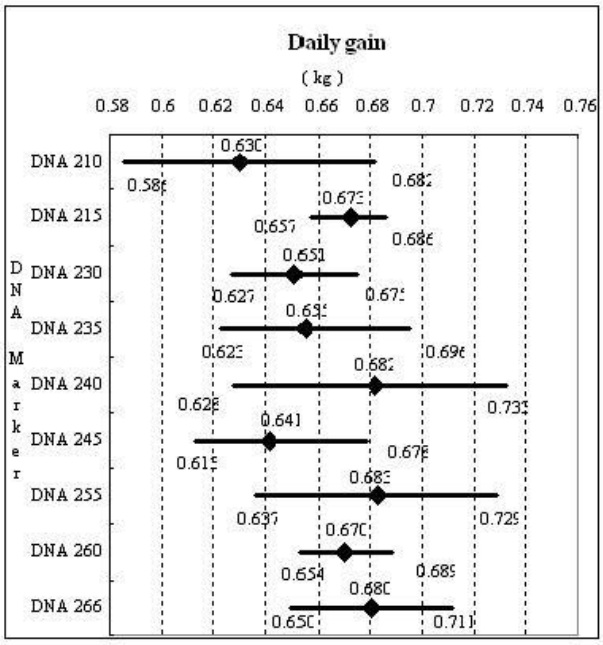

Figure 3. Bootstrap confidence intervals for daily gain.

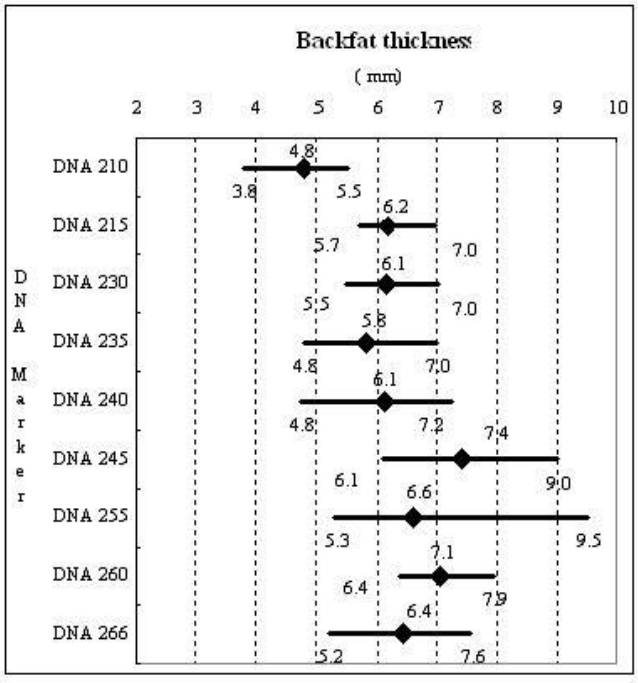

Figure 4. Bootstrap confidence intervals for backfat thickness.

have only one good marker, 266 bp, for m. longissimus dorsi area and carcass weight. But marker 210 bp is a poor marker for carcass weight (figure 6) and daily gain (figure 3). This means that markers 235 bp and 266 bp are good both for the *K*-means clustering method and for bootstrap intervals.

Summary and conclusions

LOD scores related to marbling scores and the permutation test have been applied to detect QTL. We obtained significance for loci BM3026, BMS690, ILSTS035, BM4311, BMS511 and BMC4203, but not for AFR227. We selected microsatellite locus ILSTS035 on chromosome 6 for further analysis. *K*-means clustering analysis of nine markers in ILSTS035 and five traits resulted in three cluster groups. DNA markers 210, 235 and 266 bp were selected as being the most useful in ILSTS035. Although marker 210 bp appeared to be important for backfat and

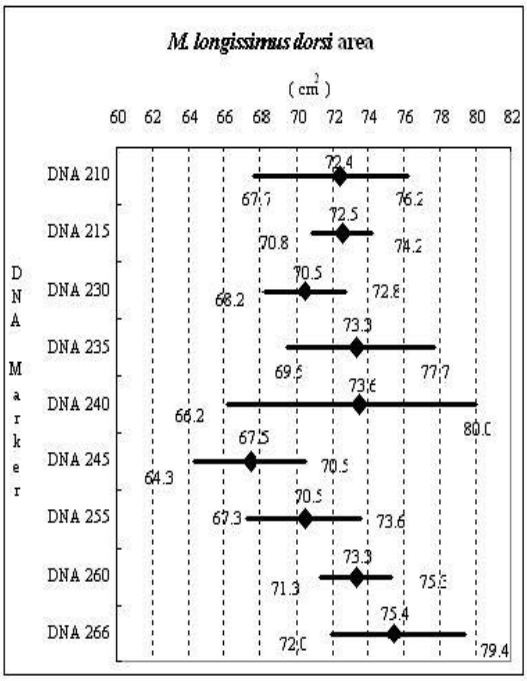

Figure 5. Bootstrap confidence intervals for m. longissimus dorsi area.

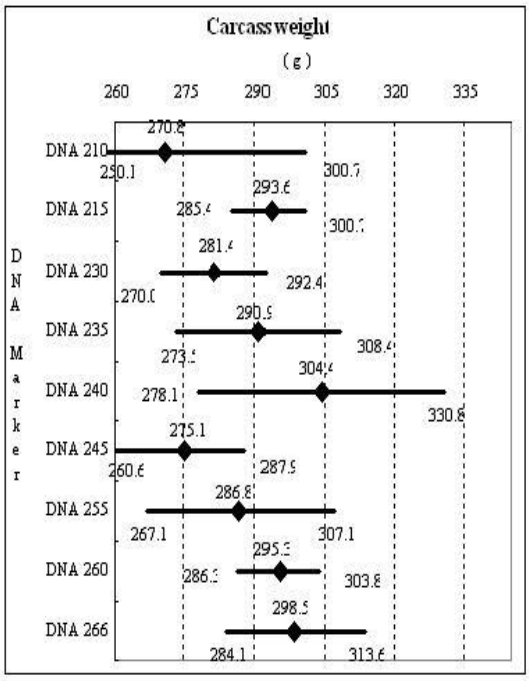

Figure 6. Bootstrap confidence intervals for carcass weight.

marker 235 bp for marbling, the individuals with these markers in our study are only 10 and 15, which may be insufficient for the conclusion. Therefore we applied the bootstrap test to calculate confidence intervals for traits. Marker 210 bp was shown to be a poor marker for carcass weight and daily gain. We conclude that the major markers of ILSTS035 locus on chromosome 6 of Hanwoo cattle are markers 235 bp and 266 bp.

Acknowledgement

This work was supported by the Korea Research Foundation grant KRF-2001-005-G20008.

References

Chotai J. 1984 On the lod score method in linkage analysis. *Ann. Hum. Genet.* **48**, 359–378.

Churchill G. A. and Doerge R. W. 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Efron B. 1987 Better bootstrap confidence intervals. *J. Am. Stat. Assoc*. **82**, 171–199.

Good P. 1994 *Permutation test: a practical guide to resampling for testing hypotheses.* Springer, New York.

Hirano T., Kobayashi N., Nakamaru T., Hara K. and Sugimoto Y. 1998 Linkage analysis of meat quality in Wagyu. 26th International Conference on Animal Genetics, Auckland, New Zealand: E019.

Kim J. W., Jang T. K., Park Y. A. and Yeo J. S. 2000 Linkage mapping of chromosome 6 in the Korean cattle (Hanwoo). *Asian-Aust. J. Anim. Sci.***13** (suppl.), 235.

Kim J. W., Park S. I. and Yeo J. S. 2003a Linkage mapping and QTL on chromosome 6 in Hanwoo (Korean cattle). *Asian-Aust. J. Anim. Sci*. **16**, 1402–1405.

Kim M. J., Lee J. Y., Yeo J. S., Lee Y. W. and Joe Y. J. 2003b A major DNA mining of BM4311 in Hanwoo. Proceedings of the Spring Conference, Cheju National University, 305–311.

Knott S. A. and Haley C. S. 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genet. Res*. **60**, 139–151.

Lander E. and Botstein D. 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

McPherron A. C. and Lee S. J. 1997 Double muscling in cattle due to mutations in the myostatin gene. *Proc. Natl. Acad. Sci. USA* **23**, 12457–12461.

MacQueen J. B. 1967 Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.* **1**, 281–297. University of California Press, Berkeley.

Visscher P., Thompson R. and Haley C. S. 1996 Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013–1020.

Weller J. I. 1986 Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**, 627–640.

Received 30 April 2004; in revised form 4 October 2004