Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 4, pp. 1103 ~ 1113

Bootstrapping and DNA marker Mining of BMS941 microsatellite locus in Hanwoo chromosome 17

Jea-Young Lee[1] · Jung-Hwan Bae[2] · Jung-Sou Yeo[3]

Abstract

LOD scores and a permutation test for detecting and locating quantitative trait loci(QTL) from the Hanwoo economic trait have been described and we selected a considerable major BMS941 locus. K-means clustering analysis of eight markers in BMS941 and four traits resulted in three cluster groups. Finally, we applied the bootstrap test method to calculate confidence intervals for finding major DNA markers. We conclude that the major markers of BMS941 locus in Hanwoo chromosome 17 are markers 85bp and 105bp.

Keywords : Bootstrap Method, K-means Clustering, LOD Score, Permutation Test, QTL.

1. Introduction

Problems detecting and locating quantitative trait loci (QTL) have received considerable attention over the past several years. A variety of methods have been developed to analyze quantitative-trait data (Weller 1986, Lander and Bostein 1989, Churchill and Deorge 1994). Many research groups(Hirano et al. 1998; Kim et al. 2000, 2003a,b; Yeo et al. 2004) have intensively analyzed linkage between markers and traits to identify the 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 the results of particular genes (McPherron and Lee, 1997). Such identification of genes

1) Professor, Department of Statistics, Yeungnam University, Gyeongsan 712-749, KOREA. Correspondence : jlee@yu.ac.kr
2) Graduate, Department of Biotechnology, Yeungnam University, Gyeongsan 712-749, KOREA.
3) Professor, Department of Biotechnology, Yeungnam University, Gyeongsan 712-749, KOREA.

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 to all of these methods is the difficulty in determining appropriate significance thresholds (critical value) 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. In 1994, Churchill and Doerge proposed permutation tests to detect the QTL effect 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 don't have theoretical significant levels (critical value or p-value). Ten thousand repetitions of the permutation process were used for critical value. This locus includes eight genes: DNA marker 80bp, 83bp, 85bp, 90bp, 97bp, 100bp, 103bp and 105bp. Next, the relations between DNA markers and the economic trait were identified by K-means clustering analysis. Finally, we applied the bootstrap test (Efron 1987; Visscher etal.,1996) to calculate confidence intervals of QTL locations for traits. The number of bootstrap samples for each DNA was 1,000 and 95% confidence intervals were calculated for economically important traits.

2. Materials and Bootstrapping (BCa (Bias-corrected and accelerated)) Analysis

Animals and Traits

Two hundred and sixty nine steers from 36 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 of 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 and K-means Clustering Methods

A permutation test 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 related to economic traits 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 procedures (Good, 1994; p20) for a permutation test. After the permutation test, we needed to identify the major DNA marker mining in BMS941 based on economically important traits such meat quality and carcass length (Lee and Lee, 2005).

Similarly, K-means clustering analysis applied to four traits, which was suggested by MacQueen(1967), is a non-hierarchical clustering technique. The results have been obtained in table 2 through table 4 with figure 2 (Lee and Lee, 2005).

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 $\hat{\theta}-\theta$ is the distribution determined by generating $\hat{\theta}$ values which are determined by sampling independently with the replacement form empirical distribution, F_n. The bootstrap estimate of the standard error of $\hat{\theta}$ then becomes the standard deviation of the bootstrap distribution for $\hat{\theta}-\theta$.

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 $\hat{\theta}$. The basic idea behind the bootstrap is that the variability of θ^* around $\hat{\theta}$ will be similar to the variability of $\hat{\theta}$ around σ. 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 θ^*, the value of $\hat{\theta}$ 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 θ^*.Let $\widehat{\theta^*_{(\alpha)}}$ indicate the 100×αth percentile of B=1000 bootstrap replications $\widehat{\theta^*_{(1)}}, \widehat{\theta^*_{(2)}}, \ldots, \widehat{\theta^*_{(B=1000)}}$.

Step 4: The BCa interval endpoints are also given by percentiles of the bootstrap distribution. The percentiles used however, depend on two numbers, $\hat{\alpha}$ (acceleration) and Z_0 (bias-correction).

The BCa interval of intended coverage 1-2α is given by BCa ;

$$(\widehat{\theta_{lo}} , \widehat{\theta_{up}}) = (\widehat{\theta^{*(\alpha 1)}} , \widehat{\theta^{*(\alpha 2)}})$$

where $\alpha_1 = \Phi\{\hat{Z}_0 + (\hat{Z}_0 + Z^{(\alpha)})/[1-\hat{\alpha}(\hat{X}_0 + Z^{(\alpha)})]\}$,

$\alpha_2 = \Phi\{\hat{Z}_0 + (\hat{Z}_0 + Z^{(1-\alpha)})/[1-\hat{\alpha}(\hat{X}_0 + Z^{(1-\alpha)})]\}$

Here $\Phi(\cdot)$ is the standard normal cumulative distribution function. $Z^{(\alpha)}$is the 100th percentile point of standard normal distribution. If $\hat{\alpha}$ and $\hat{Z}_0$ equal zero, then the BCa interval is the same as the percentile interval If $\hat{\alpha}$ and $\hat{Z}_0$ are not equal to zero, then the BCa interval endpoints change. Bias-correction $\hat{Z}_0$ is obtained from

$$\hat{Z}_0 = \Phi^{-1}\left[\frac{\sum_{b=1}^{n} I(\hat{\theta^*}(b) < \hat{\theta})}{B}\right]$$

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

3. Results

QTL Methodology

LOD scores and the permutation test for detecting and locating quantitative trait loci (QTL) from the Hanwoo economic traits 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 Deorge, 1994). An empirical 100(1-P) percentile obtained by 10,000 repetition 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, BMS8125, BMS499, BMS941, BMS1167 and HUJ223 are significant statistically, but other loci are not significant levels of P. In particular, BMS941, BMS499 and BMS1167 were demonstrated to be the best. The present work was an attempt at DNA marker mining of BMS941 microsatellite locus only in Hanwoo chromosome 17.

Table 1. QTL and permutation test of Hanwoo chromosome 17 based on economic traits

Loci	Economic Traits							
	Marbling score		Daily gain		Backfat thickness		M. Longissimus dorsi area	
	Lod Score (P-value)	Ratio of QTL variation (%)	Lod Score (P-value)	Ratio of QTL variation (%)	Lod Score (P-value)	Ratio of QTL variation(%)	Lod Score (P-value)	Ratio of QTL variation (%)
BMS 8125	3.38 (0.000)	6.22	5.28 (0.000)	9.43	2.01 (0.004)	7.54	9.24 (0.000)	15.33
BMS 1825	4.67 (0.000)	8.62	1.43 (0.072)	2.83	2.27 (0.028)	6.57	5.03 (0.000)	9.22
BMS 499	3.98 (0.000)	7.39	3.19 (0.000)	6.02	4.6 (0.000)	10.39	9.27 (0.000)	15.65
BMS 941	**3.75 (0.000)**	**7.29**	**5.7 (0.000)**	**10.72**	**4.07 (0.000)**	**4.96**	**5.11 (0.000)**	**9.67**
BMS 1101	3.15 (0.000)	5.89	2.04 (0.119)	3.92	4.28 (0.000)	6.48	6.02 (0.000)	10.68
BMS 1879	3.49 (0.000)	6.76	4.98 (0.000)	9.4	2.52 (0.015)	3.02	5.55 (0.000)	10.33
BMS 1167	4.71 (0.000)	8.8	6.14 (0.000)	11.22	5.7 (0.000)	9.96	4.98 (0.000)	9.26
HUJ2 23	3.58 (0.000)	6.46	2.3 (0.002)	4.26	5.04 (0.000)	6.25	2.98 (0.003)	5.44
CSS M033	3.79 (0.000)	7.08	3.39 (0.000)	6.42	1.59 (0.012)	3.41	2.27 (0.011)	4.38

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

K-means clustering and results

Two hundred and sixty nine steers from Hanwoo Improvement Center, National Agricultural Cooperation Federation, Korea were used for the analysis. We

analyzed the BMS941 micro locus in chromosome 17. Eight DNA markers were obtained, including 80, 83, 85 bp etc., as well as data on four economic traits namely marbling score, daily gain, backfat thickness, and M. longissimus dorsi area.

The K-means clustering analysis method applied to the four traits and eight DNA markers resulted in three cluster groups (table 2, 3 and figure. 1). From table 2, we can conclude that cluster 1 is a useful group for backfat thickness (high value=0.439), cluster 2 is a useful group for marbling score (high value=1.219), and cluster 3 is a useful group for daily gain and the M. longissimus dorsi area.

Table 2. K-means clustering analysis for economic traits

	Economic Traits			
	Marbling score	Daily gain	Backfat thickness	*M. Longissimus dorsi* area
Cluster 1(98)	-0.677	-0.822	**0.439**	-0.665
Cluster 2(87)	**1.219**	0.159	-0.173	0.241
Cluster 3(84)	-0.473	**0.797**	-0.333	**0.526**

( ) : total number of individuals

Table 3. Standardized mean results of five traits and DNA markers of BMS941

Economic Traits	DNA marker							
	80bp (8)	83bp (100)	85bp (138)	90bp (53)	97bp (14)	100bp (78)	103bp (64)	105bp (19)
Marbling score	-0.331	-0.0186	**0.1392**	-0.0879	-0.0489	-0.0718	-0.0178	**0.1576**
Daily gain	-0.8262	0.0131	**0.2013**	-0.2079	-0.4783	-0.0391	-0.0584	**0.1381**
Backfat thickness	**0.2639**	0.064	0.0121	**0.1763**	**0.3935**	-0.0896	-0.1199	-0.1194
M. Longissimus dorsi area	-0.5599	0.0186	**0.1977**	-0.2683	-0.3046	-0.1105	0.0877	-0.0643

( ) : total number of individuals

Figure1 shows that cluster 1 has a great proportion of DNA markers 90 and 103bp, cluster 2 has a great proportion of 80, 83 and 105bp, and cluster 3 has a great proportion of 97. Marker 80bp however has very few individuals (n=8) and it may not sufficient for drawing conclusions.

Similarly, we recorded standardized mean results for the four economic traits, compared with DNA markers in table 3. DNA marker 80 and 105bp present a higher marbling score, marker 80, 90 and 97bp present higher backfat thickness. DNA marker 85 and 105bp present higher M. longissimus dorsi area or daily gain.

A summary of the results is given in table 4. Marker 80 and 90bp are useful for backfat thickness; marker 105bp for marbling score and daily gain. Although markers 80 and 97bp are important for backfat, the number of the individuals are only 8 and 14, which may be insufficient for the conclusion. Therefore, we decided to try bootstrap testing.

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

Cluster group	Mean Result	K-means Mining
Backfat thickness(Cluster 1)	80, **90** and 97bp	**90** and 103bp
Marbling score(Cluster 2)	**85** and **105**bp	80, 83, **85** and **105**bp
Daily gain or M. Longissimus dorsi area (Cluster 3)	85 and **105**bp	97 and **105**bp

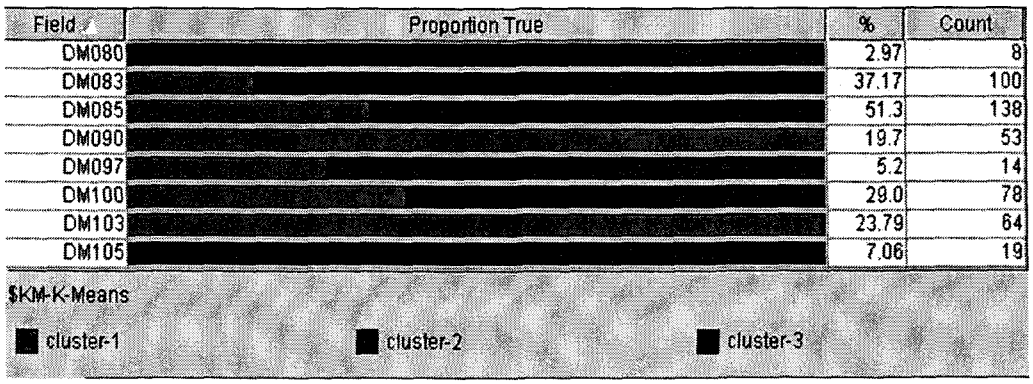

Fig. 1. Clustering proportional comparisons analysis for DNA markers(DM).

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 1,000 and 95% confidence intervals of bootstrap testing were calculated for the four traits, i.e. marbling score, daily gain, backfat thickness and M. longissimus dorsi area (figures 2 through 6).

Figure 2 shows that marker 85 and 105bp have a better marbling intervals (7.1667 ~ 8.5797) and (5.9474 ~ 10), and means 7.8696 and 7.9474 respectively than

others. In figure 3, we have markers 85 and 105bp especially good confidence intervals for daily gain. Figure 4 shows that marker 97bp has a lower backfat thickness confidence interval (4.7143 ~ 6.6429) which is good but too wide confidence interval. In figure 5, we have good DNA markers 85 and 103bp for M. longissimus dorsi area. But marker 103bp showed that it is a very bad influence marker for backfat thickness (in figure 4). This means that marker 85bp and 105bp are good both for the K-means clustering method and for bootstrap intervals.

4. Conclusions

LOD scores related to marbling scores and the permutation test have been applied for the purpose of detecting QTL. We obtained significance for loci BMS8125, BMS499, BMS941, BMS1167 and HUJ223, but not others. We selected microsatellite locus BMS 941 for further analysis. K-means clustering analysis of eight markers in BMS941 and four traits resulted in three cluster groups. DNA maker 80, 85, 97 and 105bp were selected as being the most useful genes in the BMS941. Although marker 80bp and marker 97bp appeared to be important for backfat, the individuals with these markers our study are only 8 and 14 respectively, which may be insufficient for the conclusion. Therefore we applied the bootstrap test to calculate confidence intervals for traits. DNA marker 80bp and marker 97 showed to be a bad influence marker for daily gain and M. longissimus dorsi area. We recommend that the major markers of BMS941 locus in Hanwoo chromosome 17 are markers 85bp and 105bp.

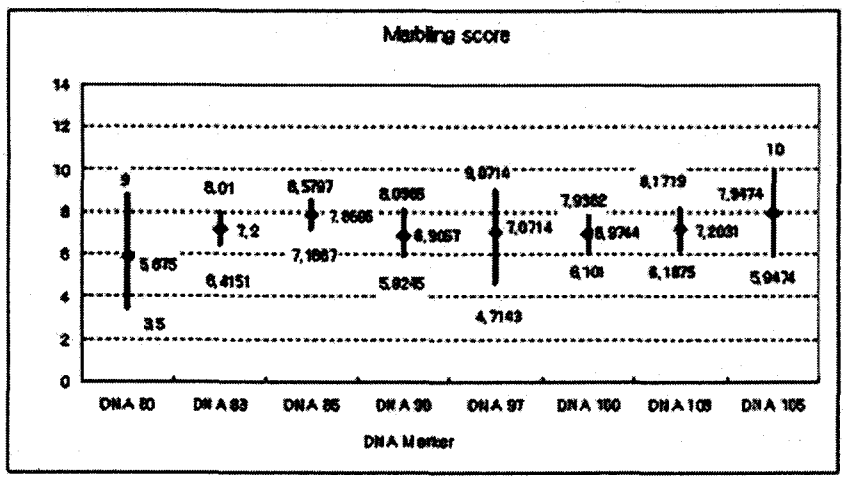

Fig. 2. Bootstrap confidence intervals of BMS941 for marbling score

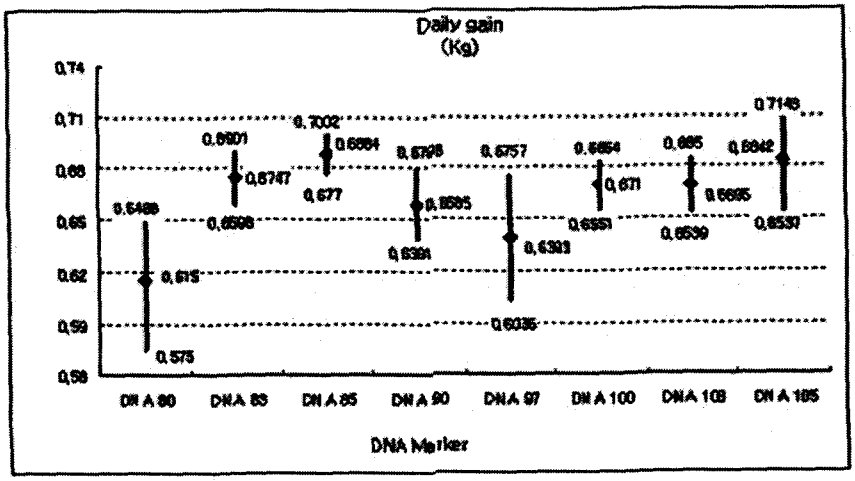

Fig. 3. Bootstrap confidence intervals of BMS941 for daily gain

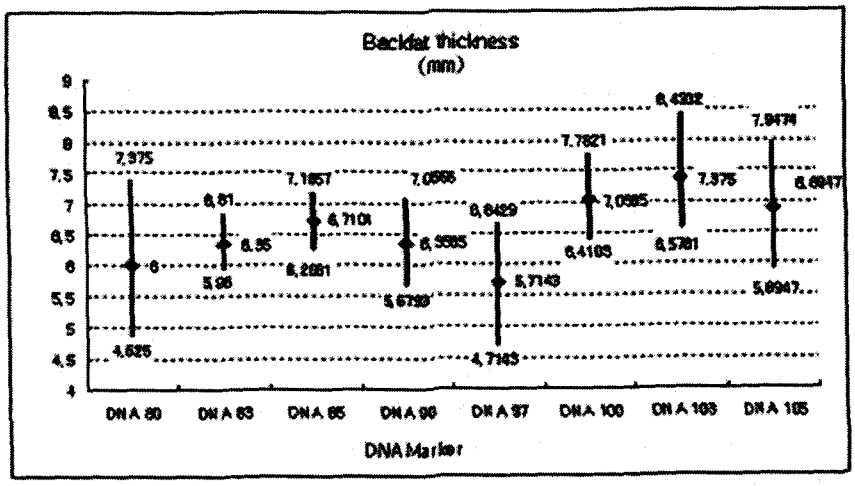

Fig. 4. Bootstrap confidence intervals of BMS941 for backfat thickness

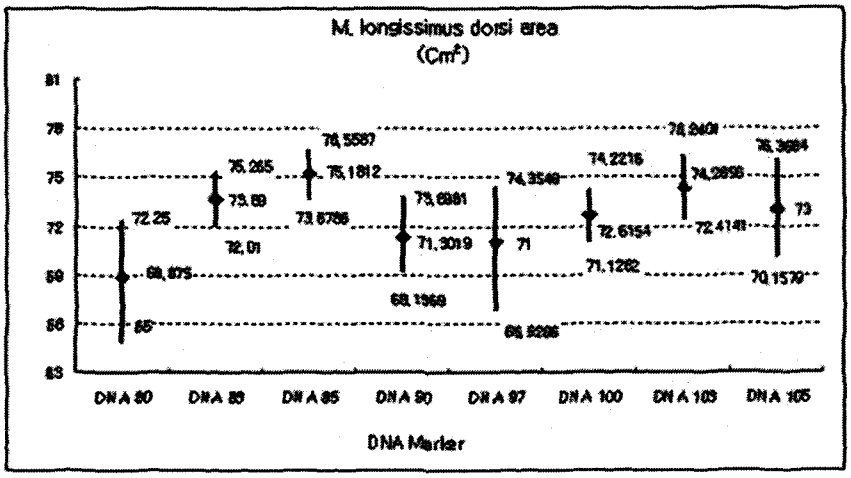

Fig. 5. Bootstrap confidence intervals of BMS941 for M. Longissimus dorsi area

References

1. Chotai, J. (1984). On the lod score method in linkage analysis. *Annals of Human Genetics* 48:359-378.
2. Churchill, G.A. and Doerge, R. W. (1994). Empirical Threshold Values for Quantitative Trait Mapping, *Genetics* 138:963-971.
3. Efron, B. (1987). Better bootstrap confidence intervals. *J. of Amer. Stat. Assoc.* 82: 171-199.
4. Good, P. (1994). *Permutation Test : A peactical Guide to resampling for testing Hypothesis*, Spring-Verlag, New York.
5. Hirano, T., Kobayashi, N., Nakamaru, T., Hara K. and Sugimoto, Y. (1998). Linkage analysis of meat quality in Wagyu. *26th International onference on Animal Genetics*, Auckland, New Zealand:E019.
6. Kim, J. W., Park, S. I. and Yeo, J. S. (2003a). Linkage Mapping and QTL on Chromosome 6 in Hanwoo (Korean Cattle). *Asian-Australasian Journal of Animal Sciences* 16(10):1402-1405.
7. 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 Univ.305-311.
8. Kim, J. W., Jang, T. K., Park, Y. A. and Yeo, J. S. (2000). Linkage mapping of chromosome 6 in the Korean Cattle(Hanwoo). 13(Suppl.): *Asian-Australasian Journal of Animal Sciences* 235.
9. Knott, S. A. and Haley C. S. (1992). Aspects of maximum likelihood method methods for the mapping of quantitative trait loci in line crosses. *GeneticalResearch* 60: 139-151.

10. Lander, E. and Bostein, D. (1989). Mapping mendelian factors underlying quantative traits using RFLP linkage maps. *Genetics*121;185-199.
11. Lee, J. and Lee, Y. (2005). A major DNA marker mining BMS941 microsatellite locus in Hanwoo chromosome 17. *J. Korean Data & Inform.* 16-4; 913-921
12. McPherron, A. C and Lee, S. J. (1997). Double muscling in cattle due to mutations in the myostatin gene. *Proceeding of National Academy of Sciences* USA 23, 12457-12461.
13. MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposiumon Mathematical Statisticsand Probability*, 1, Berkeley, California: University of California Press, 281-297.
14. Visscher, Peter, Robin, Thompson and Haley, Chris. (1996). Confidence intervals in QTL mapping by bootstrapping. *Genetics*143;1013-1020.
15. Weller, J. I. (1986). Maximum likelihood techniques for the mapping and analysis quantitative trait loci with the aid of genetic markers. *Biometrics*42:627-640.
16. Yeo, J. S., Lee, J. Y. and Kim, J. W. (2004). DNA marker mining of ILSTS035 microsatellite locus on chromosome 6 of Hanwoo cattle. *Journal of Genetics*. 83; 245-250.

[received date : Aug. 2007, accepted date : Oct. 2007]