Asian-Aust. J. Anim. Sci.
Vol. 21, No. 6 : 784 - 788
June 2008

www.ajas.info

Multifactor Dimensionality Reduction (MDR) Analysis to Detect Single Nucleotide Polymorphisms Associated with a Carcass Trait in a Hanwoo Population

Jea-Young Lee, Jae-Chul Kwon and Jong-Joo Kim[1,]*
Department of Statistics, Yeungnam University, Gyeongsan, Korea

ABSTRACT : Studies to detect genes responsible for economic traits in farm animals have been performed using parametric linear models. A non-parametric, model-free approach using the 'expanded multifactor-dimensionality reduction (MDR) method' considering high dimensionalities of interaction effects between multiple single nucleotide polymorphisms (SNPs), was applied to identify interaction effects of SNPs responsible for carcass traits in a Hanwoo beef cattle population. Data were obtained from the Hanwoo Improvement Center, National Agricultural Cooperation Federation, Korea, and comprised 299 steers from 16 paternal half-sib proven sires that were delivered in Namwon or Daegwanryong livestock testing stations between spring of 2002 and fall of 2003. For each steer at approximately 722 days of age, the *Longssimus dorsi* muscle area (LMA) was measured after slaughter. Three functional SNPs (19_1, 18_4, 28_2) near the microsatellite marker ILSTS035 on BTA6, around which the QTL for meat quality were previously detected, were assessed. Application of the expanded MDR method revealed the best model with an interaction effect between the SNPs 19_1 and 28_2, while only one main effect of SNP19_1 was statistically significant for LMA ($p<0.01$) under a general linear mixed model. Our results suggest that the expanded MDR method better identifies interaction effects between multiple genes that are related to polygenic traits, and that the method is an alternative to the current model choices to find associations of multiple functional SNPs and/or their interaction effects with economic traits in livestock populations. (**Key Words** : Multifactor-dimensionality Reduction (MDR), SNP, Hanwoo, Association Study)

INTRODUCTION

Detection of genes responsible for economic traits in farm animals has been widely practiced. Most traits of economic importance in livestock species are multi-factorial, *i.e.*, influenced by multiple genes and their interactions with environmental factors. Generally, models used to test the effects of genes on traits have been based on parametric methods, such as general linear models or the Animal model (Henderson, 1976). However, when multiple factors, *e.g.*, multiple genes and their interaction effects, are taken into account, model building may be cumbersome and over-parameterization problems can arise. As an option for efficiently detecting multiple genes and their interaction effects, a multifactor dimensionality reduction (MDR) method was introduced (Ritchie et al., 2001; Chung et al., 2006), which was designed to handle high-order dimensional data and to uncover complex relationships without relying on specified models fitting multiple genes' interactions (Bastione et al., 2004).

Most quantitative trait loci (QTL) studies have been performed using experimental crosses between different breeds or commercial populations comprising large paternal half-sib families (Kim and Farnir, 2006; Kim et al., 2007; Lee et al., 2007). Results of primary genome scans, *i.e.*, detection of QTL, are followed by candidate gene studies to find functional single nucleotide polymorphisms (SNPs) around the QTL region that are associated with the production trait of interest. As more SNP information is available due to increasing numbers of SNPs in livestock genome DBs (www.animalgenome.org) as well as in human and other mammalian genomes, selection of multiple genes around QTL, *i.e.*, functional SNPs, is a better option for simultaneous identification of several SNPs using high-throughput tools such as DNA chips (Barendse et al., 2007).

We herein report association studies using multiple SNPs by applying a MDR method to test main and

* Corresponding Author: Jong-Joo Kim. Tel: +82-53-810-3027, Fax: +82-53-810-4769, E-mail: kimjj@yumail.ac.kr
[1] School of Biotechnology, Yeungnam University, Gyeongsan, Gyeongbuk, 712-749, Korea.
Received November 6, 2007; Accepted January 9, 2008

interaction effects of multiple SNPs on meat quality around the QTL on BTA6, which is located near the ILST035 microsatellite regionin Hanwoo cattle (Yeo et al., 2004). The results were also compared with results using parametric linear models.

MATERIALS AND METHODS

Animals, phenotypes and genetic markers

Data were obtained from the Hanwoo Improvement Center, National Agricultural Cooperation Federation, Korea, and comprised 229 steers that were born at Namwon or Daegwanryong livestock testing station from 16 paternal half-sib proven sires between spring of 2002 and fall of 2003. For each steer at approximately 722 days of age, the *Longssimus dorsi* muscle area (LMA) was measured after slaughter according to the standards of the Korean Animal Products Grading Service. Six functional SNPs were selected that were located near the microsatellite marker, ILSTS035 on BTA6, around which the QTL for meat quality were detected in our previous study (Yeo et al., 2004). To determine whether the SNPs were independently distributed, linkage disequilibrium was tested between pairs of SNPs. There were complete linkage disequilibria between two SNPs and between another three SNPs, which left three independent SNPs (19_1, 18_4, 28_2) for this study (results not shown).

Statistical analysis using general linear model

The general linear model to analyze the phenotype was:

$$Y_{ijklmn} = \mu+C_i+S_j+\beta X_{age}+M1_k+M2_l+M3_m+M1M2_{kl} +M1M3_{km}+M2M3_{lm}+M1M2M3_{klm}+\varepsilon_{ijklmn}$$

Where, Y_{ijklmn} is an observed phenotype, μ is the overall mean, C_i is the i^{th} contemporary group (i = 1 to 8), S_j is the j^{th} sire's random effect (j = 1 to 16), β is a linear effect of the steer's age, Mi_k is the k^{th} genotype effect of a marker (k = 1, 2, 3) for SNP i, MiMj is an interaction term between markers i and j, M1M2M3 is a three-way interaction term for the three markers and ε_{ijklmn} is random error. The contemporary group was defined as a group of individuals fed in the same pen and slaughtered on the same date. The analyses were performed using MIXED procedure of SAS v.9.1.

MDR analysis

Multifactor-dimensionality reduction (MDR) method is non-parametric and model-free, and was initially implemented in case-control studies (Hahn et al., 2003). For application to continuous data, the CART (classification and regression tree) algorithm was developed and combined into the MDR method (Paolo, 2003). The expanded MDR method involves classification into two groups using a regression tree, *i.e.*, high and low for the phenotypes, and impurity in the group can be defined as:

$$I(g) = \frac{\sum_{j=1}^{N_g}(y_{gj} - \bar{y}_g)^2}{N_g}$$

Where, $\bar{y}_g$ is a mean value for the node (group) and y_{gj} is the observation of the j^{th} individual in a total of N_g individuals of the group (g = high or low). Each individual is assigned to a cell (*e.g.*, if there are two SNPs with three genotypes per SNP, then there are 9 possible cells), and then each cell can be defined as high or low group. The expected numbers of the high and low group are:

$$N_{high} = \sum_{i=1}^{n} I_{high}(i) \cdot N_i$$

$$N_{low} = \sum_{i=1}^{n} I_{low}(i) \cdot N_i$$

Where n is the number of total cells and $I_g(i)$ is a indicator function where

$$I_{high}(i) = \begin{cases} 1 & i(cell) \in high\ \ group \\ 0 & o.w \end{cases}$$

$$I_{low}(i) = \begin{cases} 1 & i(cell) \in low\ \ group \\ 0 & o.w \end{cases}$$

Then the average squared error (ASE) is:

$$ASE = S_{high}/N_{high}+S_{low}/N_{low}$$

where $S_{high} = \sum_{i=1}^{n} I_{high}(i) \frac{\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_i)^2}{N_i}$

is the sum of squared errors in the high group,

$$S_{low} = \sum_{i=1}^{n} I_{low}(i) \frac{\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_i)^2}{N_i}$$

is the sum of squared errors in the low group, and Y_{ij} is the j^{th} individual phenotype in the i^{th} cell.

The procedures involved in the expanded MDR method are displayed in Figure 1 and summarized in the following steps:

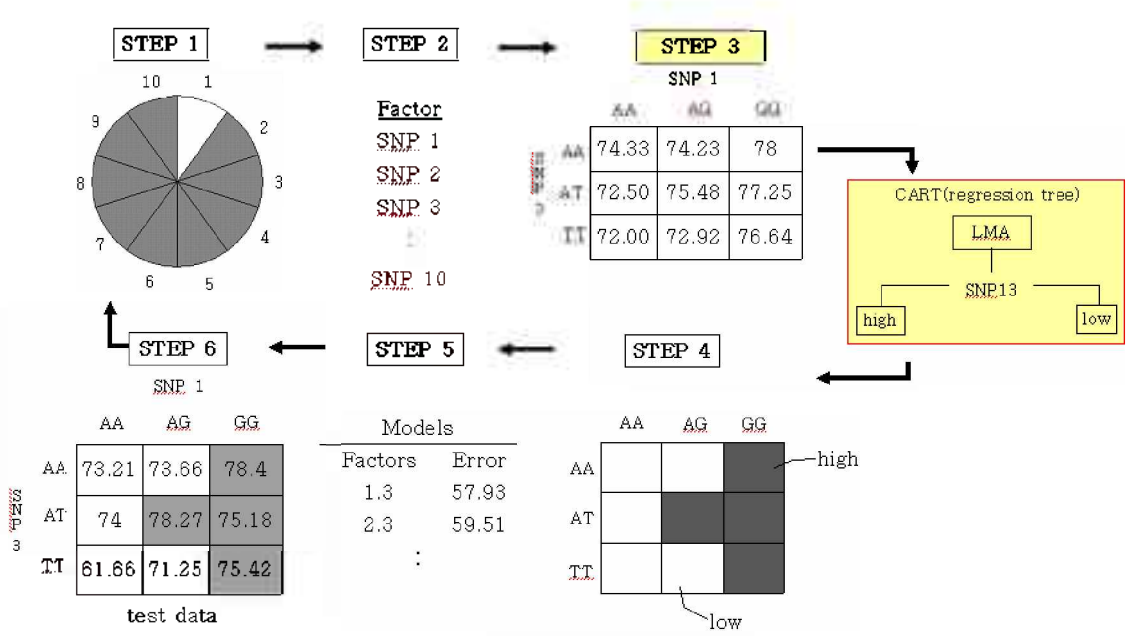

Figure 1. The expanded MDR method procedures. Steps 1-6 are repeated for each possible cross-validation interval. Numerals in cells represent the means with each multifactor combination. The darker-shaded cells in steps 4 and 6 represent high genotype combinations and the lighter-shaded cells represent low genotype combinations using the regression tree. LMA, the *Longissimus muscle dorsi area*.

Step 1. The data are randomly divided into 10 equal parts: one testing set and nine training sets as parts of the cross-validation.

Step 2. A set of n (n = 1, 2, or 3) SNPs is selected from the pool of all SNPS (= 3).

Step 3. Based on the observed level of each n, steers are partitioned into classes, referred to as cells. If n = 2, then a set of two SNPs is selected and, as one SNP has three genotypes, there are 3^2 = 9 possible cells. Phenotypic means are calculated within each cell.

Step 4. The impurity function in the CART algorithm uses the variance impurity, so that the group with the higher average value is labeled as high and the remainders are labeled as low.

Step 5. The expanded MDR model with the smallest ASE is chosen among all of the two-factor combinations (*e.g.*, SNPs 1 and 2, 1 and 3, and 2 and 3).

Step 6. In order to evaluate the predictive ability of the model, the predicted ASE (P_ASE) is estimated using 10-fold cross-validation.

These six steps were repeated for each possible combination of given n (= 1, 2, and 3). The model with the minimum predicted ASE was selected as the best-model. However, for the selected best model, statistical significance was not determined by the predicted ASE. Thus, permutation tests were performed to determine empirical significance thresholds by applying the same MDR method (Good, 1994). Before the MDR implementation, phenotypes were adjusted for the contemporary, sire and steer's age effects using residuals that were obtained after fitting the general linear model without SNP effects.

RESULTS

The effects of contemporary groups and sire effects on LMA were statistically significant ($p<0.01$). However, no interaction term between marker pairs, *i.e.*, M1M2, M1M3, M2M3, M1M2M3, was associated with the LMA variation ($p>0.3$) (results not shown). The analysis was conducted again after removing the marker interaction factors, and only one SNP (19_1) had a significant effect on LMA ($p<0.01$). For this SNP, the G allele conferred a LMA 3.1 cm^2 greater than the A allele in additive fashion, *i.e.*, with no dominance effect, explaining 4.5% of the phenotypic variation (Table 1).

Table 2 presents ASEs and P_ASEs for different combinations of SNPs that were obtained by applying the expanded MDR method to LMA analysis. Among the models with one SNP, the model with SNP19_1 had the smallest P_ASE value of 599.1. However, when considering two SNPs, the model with SNP19_1 and SNP28_2 had a the P_ASE value of 596.71, which was lower than that for the one SNP model with SNP19_1, and thus represented the best model among all combinations of SNPs. Permutation tests also revealed statistical significance (p = 0.004) for the interaction effects of

Table 1. Least squares means and standard errors (SE) of the SNP genotypes for *Longissimus muscle dorsi* area (LMA) using a general linear model*

SNP	Genotype	No. of animals	Mean±SE (cm^2)
19_1	AA	16	70.6±2.1[a]
	AG	75	73.4±1.0[a]
	GG	138	76.8±0.9
p-value			0.001
18_4	CC	45	72.2±1.4[a]
	CT	112	74.4±1.1[a]
	TT	72	74.3±1.2[a]
p-value			0.220
28_2	AA	55	73.7±1.3[a]
	AT	120	74.3±1.1[a]
	TT	57	72.9±1.3[a]
p-value			0.574

* The linear model included a fixed effect of contemporary groups, a covariate of steer's age, and a random effect of sire. In addition, three SNPs were fitted as fixed factors in the model.

[a] The same letter indicates no significant difference between the means ($p>0.05$).

SNP19_1 and SNP28_2 ($p = 0.0045$ for the one SNP model with SNP19_1).

DISCUSSION

A novel method, MDR, was applied to reduce the dimensionality caused by simultaneously fitting multiple genes and their interactions into models. In this MDR method, a CART algorithm was added to adjust for continuous properties of phenotypes. This model-free and non-parametric method produced different results when comparing with the linear mixed model. While the latter model did not detect any interaction effects between the SNPs, but detected only the main effect of one SNP, SNP19_1, the MDR method enabled the choice of the best model with an interaction effect between two SNPS, SNP19_1*SNP28_2 (Tables 1 and 2). In the linear mixed model, factors to estimate parameters are fitted in orderly form, *i.e.*, main factors that are followed by interaction effects between the main effects. That is, quadratic or higher-orderly effects are tested as conditional on the main effects. In contrast, the MDR method is free of factor-dimension orders such that a model that has an effect with the most significant contribution to phenotypic variation can be chosen first. Intrigued by the results of the MDR analysis, another linear parametric model was applied in which, among all possible SNP effects, only the interaction effect of SNP19_1*SNP28_2 was fitted without main effects of SNPs. Sum of squares explained by the interaction term was 940.7, which was greater than the value of 661.0 for SNP19_1 when fitting only main SNP effects in the model (Table 1). However, the p value for the interaction effect was not lower (0.018) than that ($p = 0.002$) under the model with main SNP effects, because the p value is a function of the sum of squares and degrees of freedom between numerator and denominator in the *F*-statistics as well as factors fitted in the model.

Table 2. Average squared error (ASE) and predicted ASE (P_ASE) for different numbers of SNPs using the expanded MDR method*

No. of SNPs	SNP	ASE	P_ASE
1	19_1	588.4	599.1
	18_4	606.2	615.6
	28_2	602.8	604.8
2	19_1*18_4	583.8	606.8
	19_1*28_2	579.3	596.7
	18_4*28_2	591.5	615.0
3	19_1*18_4*28_2	565.3	633.0

* The method was non-parametric and model-free, to reduce multifactor-dimensionalities of interaction terms between multiple SNPs. ASE or P_ASE indicates how much error occurred in determined groups, defined as the actual numbers of phenotype-high or -low groups compared to expected numbers of (phenotype) high or low groups.

The number of SNPs used in this study to evaluate effects of high-order interactions under the parametric and non-parametric models was limited in comparison to other MDR reports, in which more than 10 genes or SNPs were tested in human multi-factorial diseases (Bastone et al., 2004; Cho et al., 2004). However, the MDR method applied in this study allowed detection of an interaction effect of two SNPs, while the parametric linear model did not. This result suggests that the expanded MDR method is an alternative to model choices that can find associations of multiple functional SNPs and/or their interaction effects with economic traits in livestock. Indeed, most of these traits are influenced by multiple genes with environmental interactions, which may be strongly affected by interactions among genes (Carlborg and Haley, 2004).

One disadvantage of using the expanded MDR method is that this MDR method requires demanding computational analysis when applied at genome-wide level, *e.g.*, using 10 K or 50 K DNA chips (Barendse et al., 2007). However, application of the method with moderate numbers of genes or SNPs, *e.g.*, around several QTL regions that were previously detected, may be a good alternative for better identifying interaction effects between genes that are related to polygenic traits.

ACKNOWLEDGMENT

Lee, J.-Y. was supported by the Yeungnam University research grants in 2007.

REFERENCES

Barendse, W., A. Reverter, R. J. Bunch, B. E. Harrison, W. Barris and M. B. Thomas. 2007. A validated whole-genome

association study of efficient food conversion in cattle. Genetics 176:1893-1905.

Bastone, L., M. Reilly, D. J. Rader and A. S. Foulkes. 2004. MDR and PRP: a comparison of methods for high-order genotype-phenotype associations. Human Heredity 58:82-92.

Carlborg, O. and C. S. Haley. 2004. Epistasis: too often neglected in complex trait studies? Nat. Rev. Genet. 5:618-625.

Cho, Y. M., M. D. Ritchie, J. H. Moore, J. Y. Park, K.-U. Lee, H. D. Shin, H. K. Lee and K. S. Park. 2004. Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. Diabetologia 47:549-554.

Chung, Y. J., S. Y. Lee and T. S. Park. 2006. Multifactor dimensionality reduction in the presence of missing observations. In: Proceedings of the 2006 Spring Korean Statistical Society Conference, Pusan National University, Pusan. pp. 31-36.

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

Hahn, L. W., M. D. Ritchie and J. H. Moore. 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19:376-382.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in predicting of breeding values. Biometrics 32:69-83.

Kim, E. H., B. H. Choi, K. S. Kim, C. K. Lee, B. W. Cho, T.-H. Kim and J.-J. Kim. 2007. Detection of Mendelian and parent-of-origin quantitative trait loci in a cross between Korean native pig and landrace I. growth and body composition traits. Asian-Aust. J. Anim. Sci. 20:669-676.

Kim, J.-J. and F. Farnir. 2006. Evaluation of a fine-mapping method exploiting linkage disequilibrium in livestock populations: simulation study. Asian-Aust. J. Anim. Sci. 19: 1702-1705.

Lee, Y.-M., J.-H. Lee and J.-J. Kim. 2007. Evaluation of reciprocal cross design on detection and characterization of non-Mendelian QTL in F2 outbred populations: I. parent-of-origin effect. Asian-Aust. J. Anim. Sci. 20:1805-1811.

Paolo, G. 2003. Applied data mining: Statistical methods for business and industry (Statistics in practice). John Wiley & Sons, England.

Ritchie, M. D., L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl and J. H. Moore. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Human Genet. 69(1):138-147.

Yeo, J. S., J. Y. Lee and J. W. Kim. 2004. DNN marker mining of ILST035 microsatellite locus on chromosome 6 of Hanwoo cattle. J. Genet. 83:245-250.