응용통계연구 (2010)
23(4), 693–703

우수 유전자 조합 선별을 위한 통계적 상호작용 방법비교

이제영[1] · 이용원[2] · 최영진[3]

[1]영남대학교 통계학과, [2]영남대학교 통계학과, [3]영남대학교 통계학과

(2010년 2월 접수, 2010년 7월 채택)

요 약

대개 인간의 질병과 관련된 유전자나 가축의 경제적인 특성과 관련된 유전자는 주로 상호작용으로 일어난다. 유전자의 상호작용을 찾기 위한 방법으로 다양한 방법들이 제시되었다. 본 논문에서는 유전자의 상호작용 효과를 규명하기 위해 개발된 확장된 MDR방법(E-MDR)과 더미변수를 활용한 MDR방법(D-MDR), 대규모 유전자들 중에서 주요 유전자 조합을 선별하는 SNPHarvester방법을 비교하여 인간의 질병이 아닌 한우의 경제적인 특성에 적용하여 우수한 유전자 조합을 선별하고 우수 유전자형을 밝힌다.

주요용어: E-MDR, D-MDR, SNPHarvester, 유전자 조합, 유전자형.

1. 서론

광범위 유전자 연관(genome-wide association; GWA) 연구에서는 수많은 단일 염기 다형성(single nucleotide polymorphisms; SNPs)을 사용하여 인간 질병에 포함되어 있는 유전자를 식별하는데 많은 노력을 해왔다. SNP들의 상호작용은 질병이나 가축의 경제형질의 특성을 발견하는데 매우 중요한 역할을 하는데 그 이유는 인간의 질병과 관련된 유전자 혹은 가축의 경제형질 특성에 관련된 유전자가 대개 상호작용으로 존재하기 때문이다. 유전자의 상호작용을 계산하기위해 이전에는 주로 선형모형과 같은 통계모형을 사용해왔다. 그러나 유전자의 수가 많아짐에 따라 모형이 복잡해지고 해석하는 것에 어려움이 생겼다. 이러한 문제점을 해결하기 위해 다양한 통계적 방법들이 제시되어왔다. 인간의 질병에 대한 유전자의 상호작용을 찾는 방법으로 다중인자 차원 축소방법(Multifactor Demensionality Redution; MDR; Ritchie 등, 2001; Chung 등, 2005), CART(classification and regression tree)방법을 활용한 확장된 다중인자 차원 축소방법(Expanded Multifactor Dimensionality Redution; E-MDR; Lee 등, 2008), 더미변수를 활용한 다중인자 차원 축소방법(Dummy Multifactor Dimensionality Redution; D-MDR; Lee 등, 2008; Lee 등, 2009) 등이 개발되었다. 이 방법들 중 MDR방법은 사례-대조로 구성되어 있는 이분화 데이터에 적용하여 복잡한 유전자의 상호작용을 계산한다. 그러나 MDR방법은 이분화 데이터에만 적용 가능하다는 한계점을 가진다. 이를 연속형 자료에도 적용하기 위해 개발된 방법이 CART방법을 활용한 E-MDR방법과 더미변수를 활용한 D-MDR방법이다. 유전자의 상호작용을 찾기 위한 또 다른 방법으로 SNPHarvester (Yang 등, 2009)가 개발되었다. SNPHarvester는 대규모의 유전자들 가운데서 주요한 유전자 조합을 찾는 방법이다. 이 방법 역시 인간의 질병의 유무와 관련된 주요 유전자 그룹을 찾는 방법으로 제시되었기 때문에 주로 사례-대조군과 같은 이분화 된 데이터에 적용이 가능하다. 본 논문에서는 연속형 자료인 한우자료에 E-MDR, D-MDR, SNPHarvester를 적용하여 한

[1]교신저자: (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: jlee@yu.ac.kr

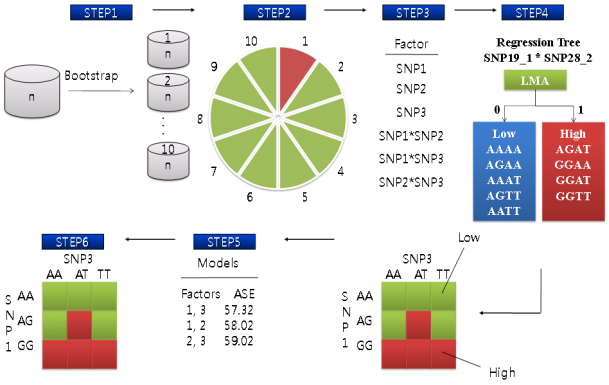

그림 **2.1.** E-MDR 방법의 일반적인 절차

우의 경제형질에 가장 많은 영향을 주는 우수한 유전자 조합을 선별하고자 한다. 세 방법을 한우의 경제형질에 모두 적용하여 공통적으로 선별된 유전자 조합을 한우의 경제형질에 영향을 미치는 우수 유전자 조합으로 본다. E-MDR, D-MDR, SNPHarvester는 유전자의 주요 상호작용을 찾는 방법으로 개발되었기 때문에 각 방법들에서 선별된 공통의 유전자가 한우의 경제형질에 영향을 주는 우수 유전자 조합으로 활용할 수 있을 것이다. 따라서 우리는 E-MDR, D-MDR, SNPHarvester방법을 소개하고 한우 자료에 각 방법을 적용하여 한우의 경제형질인 일당증체량(average daily gain; ADG), 도체중량(carcass cold weight; CWT), 근내지방도(marbling score; MS)에서 우수 유전자 조합을 선별한다. 또한 선별된 우수 유전자 조합에서 우수 유전자형(genotype)을 찾기 위해 데이터 마이닝 기법 중 하나인 CART방법을 이용한다. 본 논문의 2장에서는 우수 유전자 조합을 선별하기 위한 통계적 방법에 대해서 소개한다. 그리고 3장에서는 본 논문에서 사용된 실험자료를 살펴보고 우수 유전자 조합을 선별하기 위한 통계적 방법들을 한우 자료에 적용하여 우수 유전자 조합을 선별한다. 4장에서는 우수 유전자 조합에서 찾아진 우수 유전자형을 규명한다.

2. 우수 유전자 조합 선별을 위한 통계적 방법

2장에서는 우수 유전자 조합을 선별하기 위한 통계적 방법들에 대해서 살펴보고 각 방법들의 특징을 소개한다. 또한 이들의 결과를 검정하기 위해 사용한 순열 검정 방법도 소개한다.

2.1. CART 알고리즘을 활용한 확장된 MDR 방법

MDR (Ritchie 등, 2001)방법은 인간의 질병 유무와 같은 이분형 자료에 대한 유전자의 상호작용을 찾기 위해 제시된 방법이다. 이 방법은 사례-대조로 이루어진 이분형 데이터에만 적용이 가능하다는 한계점을 가지고 있기 때문에 Lee 등 (2008)은 데이터 마이닝 기법 중 CART 알고리즘을 이용하여 연속형 자료에서도 MDR방법을 활용할 수 있는 E-MDR (Lee 등, 2008)방법을 개발하였다. 즉, E-MDR방법은 목표변수가 연속형인 경우에 CART방법으로 이분화하여 적용함으로서 MDR방법을 사용할 수 있도록 제시된 방법이다. 그림 2.1은 연속형 데이터에 대한 E-MDR방법의 절차를 도식화한 것이다.

절차 1. 데이터를 랜덤으로 10개의 같은 크기로 나누어 9개를 훈련용 데이터 셋으로, 나머지 1개를 검정용 데이터 셋으로 정한다.

절차 2. 모든 SNP로부터 k개의 SNP조합 중 하나를 선택한다.

절차 3. 선택된 SNP조합에서 SNP의 각각 수준을 기초로 한 개체들을 다중요인 집합 또는 셀에 기술한다. 예를 들어서 $k = 2$, SNP가 3개 수준으로 되어있다면 셀의 개수는 $3^2 = 9$개이며, 각 셀에 연속형 자료의 평균값을 기술한다.

절차 4. CART방법의 불순도 함수를 사용하여 이분화 한다. 분류 결과로 나온 그룹을 평균이 높은 상위그룹과 평균이 낮은 하위그룹으로 나눈다.

절차 5. 훈련용 데이터셋에서개 SNP의 모든 조합에 대한 ASE(Average Squared Error)를 다음과 같이 계산한다.

$$\text{ASE} = \frac{S_{high}}{N_{high}} + \frac{S_{low}}{N_{low}},$$

$$S_{high} = \sum_{i=1}^{n} I_{high}(i) \sum_{j=1}^{N_i} (y_{ij} - \hat{y}_i)^2, \qquad S_{low=} \sum_{i=1}^{n} I_{low}(i) \sum_{j=1}^{N_i} (y_{ij} - \hat{y}_i)^2,$$

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

$$I_{high}(i) = \begin{cases} 1, & i(\text{cell}) \in \text{high group}, \\ 0, & \text{o·w}, \end{cases} \qquad I_{low}(i) = \begin{cases} 1, & i(\text{cell}) \in \text{low group}, \\ 0, & \text{o·w}, \end{cases}$$

(n : 범주의 수, $N_{\text{ı}} : i$ 범주의 관측치 수)

절차 6. 나머지 1/10의 검정용 데이터셋을 이용하여 P-ASE(Prediction Average Squared Error)를 구한다.

위 과정의 반복에서 나온 10개의 ASE와 P-ASE의 평균을 구해 그 값이 가장 낮은 것을 우수 모형으로 정한다 (Bastone 등, 2004). 그리고 각각의 ASE를 이용하여 10번의 반복시행을 할 때 각 시행에서 각 모형이 우수 모형으로 선택된 횟수인 CVC를 계산한다 (Chung 등, 2005). 따라서 ASE와 P-ASE의 평균이 가장 낮고 CVC가 가장 높은 값을 갖는 모형을 우수 모형으로 판단한다. 본 논문에서는 이러한 E-MDR 연구를 바탕으로 한우의 경제형질인 일당증체량, 도체중량, 근내지방도에 영향을 주는 우수 유전자 조합을 선별하고자 한다.

2.2. 더미변수를 활용한 MDR방법

사례-대조로 이루어진 이분형 데이터에만 적용 가능한 MDR방법을 연속형 데이터에 적용하기 위해 개발된 또 다른 방법으로 Lee 등 (2009)은 더미변수를 활용하는 D-MDR (Lee 등, 2008; Lee 등, 2009)방법을 개발하였다. D-MDR방법은 MDR방법에서 더미변수 회귀분석을 적용하여 연속형 자료를 이분화한 후 MDR방법을 활용할 수 있도록 제시한 기법이다. D-MDR방법은 문제를 해결하기 위해 더미변수를 활용한 회귀분석방법 (김태근, 2006)을 활용한다.

그림 2.2는 연속형 데이터에 대한 D-MDR방법의 절차를 도식화한 것이다. 더미변수를 활용해 이분화한 부분(절차 4)을 제외하고 나머지 절차는 E-MDR의 절차와 동일하다. 본 논문에서는 한우의 경제형질과 연관된 유전자 상호작용을 분석하기 위해 D-MDR방법을 적용하여 우수 유전자 조합을 선별한다.

2.3. SNPHarvester방법

유전자의 상호작용을 규명하기 위해 개발된 많은 통계적 방법들은 수많은 SNP들에 적용할 시 많은 시간과 비용이 들어간다. 이러한 단점을 해결하기 위해 SNPHarvester (Yang 등, 2009)방법이 개발되

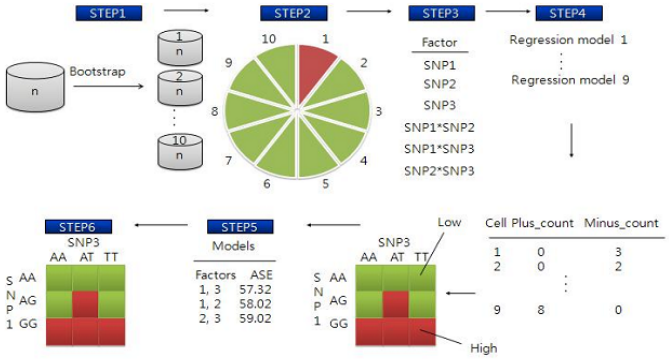

그림 **2.2.** D-MDR 방법의 일반적인 절차

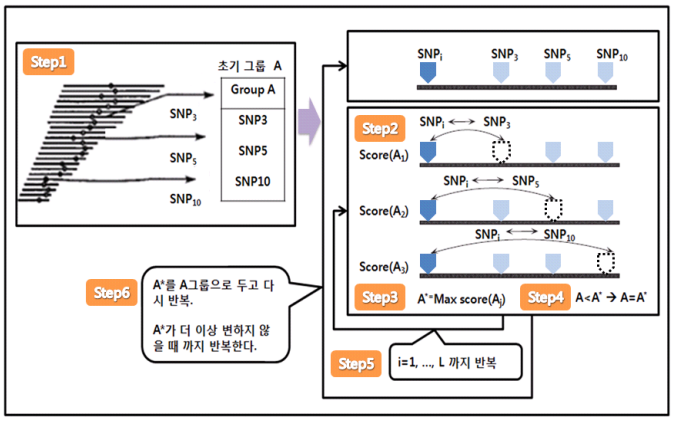

그림 **2.3.** SNPHarvester 방법의 일반적인 절차

었다. 즉, SNPHarvester는 대규모의 유전자들 가운에서 주요 유전자 조합을 선별해 내는 방법이다. 이 방법은 특성치에 연관된 SNP 그룹을 찾기 위하여 SNP들 중에서 초기에 몇 개의 SNP를 선정하여 SNP 그룹으로 결정하고 결정된 그룹에서 SNP를 하나씩 바꿔가며 스코어를 높이는 과정을 반복하는 것이다. 그림 2.3은 SNPHarvester 방법의 절차를 도식화한 것이다.

이 방법에서 스코어 함수는 MDR방법에서의 정확도, χ^2값, B-통계량 등이 사용 될 수 있다. 본 논문에서는 값을 스코어 함수로 사용한다. 또한 통계적으로 유의성을 가지는 그룹을 선별하기 위해 $\alpha = 0.05$로 정하고 초기에 선별된 그룹수인 k가 클수록 SNP 그룹 내의 유전자형의 빈도가 없거나 작아질 수 있기 때문에 $k <= \ln_3 N_d - 1$ (Yang 등, 2009)로 제한한다. 여기서 $N_d - 1$는 case의 수이다. 본 논문에서는 SNPHarvester방법을 적용하여 한우의 경제형질에 영향을 주는 우수 유전자 조합을 선별한다.

2.4. 순열 검정

E-MDR, D-MDR, SNPHarvester에 의해 선별된 유전자 조합과 유전자형이 통계적으로 유의성을 가지는지 알아보기 위해 t 검정과 순열 검정 (Good, 2000)의 p-value를 사용한다. 본 논문에서는 10,000번을 시행하여 다음의 절차에 따라 순열 검정을 하였다.

절차 1. 검정하고자 하는 가설 설정 - 각 방법에 의해 선별된 우수 유전자 조합의 유전자 형이 한우의 경제형질에 유의한 영향을 미친다.

절차 2. 통계량과 기각역 설정 - 분석에 사용할 통계량으로 'high'와 'low'그룹으로 이분화 했을 때 'high'그룹의 평균을 선택한다. 선정된 'high'그룹의 평균이 그룹간 데이터를 서로 바꾸었을 때보다 높다면 그 그룹이 경제형질에 영향력이 있다고 판단한다.

절차 3. 기존 관측치의 통계량 계산 - 선별된 'high'그룹의 평균을 계산한다.

절차 4. 관측치의 재배열과 재배열 후의 통계량 계산 - 두 그룹의 데이터를 n개 만큼 랜덤추출하여 그룹을 상호 변경한 후 각 'high'그룹의 평균을 구한다. 이 과정을 10,000번 반복한다.

절차 5. 결론 - 각 평균을 내림차순으로 정렬한 후 p-value를 구한다.

위의 순열 검정을 통해 우수 유전자 조합에서 발견한 우수 유전자형이 한우의 경제형질에 미치는 영향을 4장의 결과를 통해서 살펴본다.

3. 우수 유전자 조합을 선별하기 위한 통계적 방법 적용 및 결과

3.1. 실험자료

본 연구 데이터는 농협중앙회 한우개량사업소의 후대검정집단인 30차에서 35차 국가 후대검정우 집단 476두로 구성되어졌다. 한우의 여러 경제형질인 일당증체량(average daily gain; ADG), 도체중량(carcass cold weight; CWT), 근내지방도(marbling score; MS)는 모든 F1 자손으로부터 수집되어졌고 한국축산물등급판정소의 규격에 따라 측정되었다. Snelling 등 (2005)이 연구한 EST-based SNP 연관지도에서 Kim 등 (2003)의 연구로 규명된 한우 염색체 6번에 위치한 후보 QTL인 ILSTS035 microsatellite 마커와 같은 거리에 있는 SNP들 중 Lee (2009)의 연구를 통하여 후보 유전자로 판단되는 LOC534614 유전자내 SNP 20개에 대하여 htSNP(haplotype-tagging SNP) 방법으로 선택된 최종 6개의 SNP들(g.4102+36T>G, g.8778G>A, g.11500-117C>G, g.32330-48A>G, g.34425+102A>T, g.66995-169insdelC)을 분석에 사용하였다. 또한 한우데이터 476두에서 결측치를 제거한 후 붓스트랩(Efron과 Tibshirani, 1993)방법을 사용하여 10배를 증폭시킨 4190두를 사용하였다. 3.2∼3.4절에서는 앞장에서 소개된 E-MDR방법과 D-MDR방법, SNPHarvester방법을 적용하여 한우의 주요 경제형질에서 우수 유전자 조합을 선별한다.

3.2. 확장된 MDR 방법(E-MDR)을 활용한 유전자 조합 선별 결과

한우의 주요 경제형질인 일당증체량, 도체중량, 근내지방도에 E-MDR방법을 적용하였다. 각 경제형질의 유전자 조합에 대하여 E-MDR 과정을 10번 반복하여 ASE(Average Squared Error)와 P-ASE(Prediction Average Squared Error)의 평균 (Bastone 등, 2004)을 바탕으로 P-값이 유의하고 동시에 ASE와 P-ASE가 낮은 상위 5개 유전자 조합을 선별한 것을 표 3.1에 나타냈다.

E-MDR방법을 각 경제형질에 적용한 결과 2개의 요인으로 결합된 SNP들 중에서 일당증체량, 도체중량, 근내지방도 모두에서 공통으로 나타난 유전자 조합 즉, 종합적인 경제형질에 관계된 우수 유전자 조합으로 g.8778G>A과 g.11500-117C>G, g.11500-117C>G과 g.32330-48A>G이 선별되었다. 또한 각 경제형질에서 3개의 요인으로 결합된 SNP들 중 종합적인 경제형질에 관계된 우수 유전자 조합은 g.8778G>A, g.11500-117C>G, g.32330-48A>G과 g.4102+36T>G, g.11500-117C>G, g.32330-48A>G으로 선별되었다. 따라서 E-MDR방법에서 한우의 종합적인 경제형질에 영향을 주는 우수 유전

표 **3.1.** E-MDR을 활용한 한우의 주요 경제형질에 대한 우수 유전자 조합 선별 결과

경제형질	요인수	유전자조합	ASE	P_ASE
일당증체량	2	g.4102+36T>G, g.11500-117C>G	0.007796	0.007696
		g.8778G>A, g.11500-117C>G	0.007846	0.007747
		g.11500-117C>G, g.32330-48A>G	0.007852	0.007768
		g.8778G>A, g.66995-169insdelC	0.007907	0.007801
		g.11500-117C>G, g.34425+102A>T	0.007909	0.007819
	3	**g.4102+36T>G, g.11500-117C>G, g.32330-48A>G**	0.007755	0.007621
		g.4102+36T>G, g.11500-117C>G, g.66995-169insdelC	0.007770	0.007685
		g.4102+36T>G, g.11500-117C>G, g.34425+102A>T	0.007772	0.007700
		g.4102+36T>G, g.8778G>A, g.11500-117C>G	0.007780	0.007679
		g.8778G>A, g.11500-117C>G, g.32330-48A>G	0.007805	0.007745
도체중량	2	**g.11500-117C>G, g.32330-48A>G**	1120.013	1118.707
		g.8778G>A, g.11500-117C>G	1132.281	1130.854
		g.32330-48A>G, g.34425+102A>T	1135.011	1133.720
		g.4102+36T>G, g.32330-48A>G	1135.057	1134.640
		g.4102+36T>G, g.8778G>A	1136.854	1135.398
	3	g.11500-117C>G, g.32330-48A>G, g.66995-169insdelC	1111.296	1109.692
		g.11500-117C>G, g.32330-48A>G, g.34425+102A>T	1118.838	1120.161
		g.8778G>A, g.11500-117C>G, g.32330-48A>G	1119.860	1120.966
		g.4102+36T>G, g.11500-117C>G, g.32330-48A>G	1120.013	1118.707
		g.8778G>A, g.11500-117C>G, g.66995-169insdelC	1120.379	1124.680
근내지방도	2	g.4102+36T>G, g.11500-117C>G	16.1373	16.3072
		g.11500-117C>G, g.32330-48A>G	16.1622	16.3315
		g.8778G>A, g.11500-117C>G	16.2162	16.4331
		g.11500-117C>G, g.34425+102A>T	16.2794	16.4872
		g.8778G>A, g.32330-48A>G	16.3377	16.5118
	3	**g.4102+36T>G, g.11500-117C>G, g.32330-48A>G**	15.9860	16.1575
		g.4102+36T>G, g.11500-117C>G, g.34425+102A>T	16.0071	16.2227
		g.11500-117C>G, g.32330-48A>G, g.66995-169insdelC	16.0591	16.2512
		g.4102+36T>G, g.11500-117C>G, g.66995-169insdelC	16.0934	16.2656
		g.8778G>A, g.11500-117C>G, g.32330-48A>G	16.0938	16.3804

자 조합은 (g.8778G>A, g.11500-117C>G), (g.11500-117C>G, g.32330-48A>G), (g.8778G>A, g.11500-117C>G, g.32330-48A>G), (g.4102+36T>G, g.11500-117C>G, g.32330-48A>G)이 선별됐다.

3.3. 더미변수를 활용한 MDR 방법(D-MDR)을 활용한 유전자 조합 선별 결과

한우의 주요 경제형질에 D-MDR방법을 적용하였다. 각 경제형질의 유전자 조합에 대하여 D-MDR 과정을 10번 반복하여 ASE와 P-ASE의 평균 (Bastone 등, 2004)을 바탕으로 P-값이 유의하고 동시에 ASE와 P-ASE가 낮은 상위 5개 유전자 조합을 선별한 것을 표 3.2에 나타냈다.

D-MDR방법을 각 경제형질에 적용한 결과 2개의 요인으로 결합된 SNP 그룹에서 종합적인 경제형질에 관한 유전자 조합은 g.8778G>A과 g.11500-117C>G, g.32330-48A>G과 g.34425+102A>T이 선별되어졌다. 또한 3개의 요인으로 결합된 SNP 그룹에서의 종합적인 경제형질에 대한 유전자 조합으로 g.8778G>A, g.11500-117C>G, g.32330-48A>G이 선별되었다. 따라서 D-MDR방법에서 한우의 종합적인 경제형질에 영향을 주는 우수 유전자 조합으로 (g.8778G>A, g.11500-117C>G), (g.32330-

표 **3.2.** D-MDR을 활용한 한우의 주요 경제형질에 대한 우수 유전자 조합 선별 결과

경제형질	요인수	유전자조합	ASE	P_ASE
일당증체량	2	g.4102+36T>G, g.11500-117C>G	0.007821	0.007775
		g.11500-117C>G, g.32330-48A>G	0.007853	0.007799
		g.8778G>A, g.11500-117C>G	0.007857	0.007809
		g.11500-117C>G, g.34425+102A>T	0.007951	0.007911
		g.32330-48A>G, g.34425+102A>T	0.007978	0.007917
	3	g.4102+36T>G, g.11500-117C>G, g.66995-169insdelC	0.007524	0.007479
		g.4102+36T>G, g.11500-117C>G, g.34425+102A>T	0.007540	0.007504
		g.4102+36T>G, g.8778G>A, g.11500-117C>G	0.007580	0.007542
		g.4102+36T>G, g.11500-117C>G, g.32330-48A>G	0.007584	0.007535
		g.8778G>A, g.11500-117C>G, g.32330-48A>G	0.007596	0.007563
도체중량	2	g.11500-117C>G, g.32330-48A>G	1120.980	1112.419
		g.8778G>A, g.11500-117C>G	1132.281	1125.725
		g.32330-48A>G, g.34425+102A>T	1135.011	1126.386
		g.4102+36T>G, g.32330-48A>G	1135.896	1126.899
		g.4102+36T>G, g.8778G>A	1136.854	1129.715
	3	g.8778G>A, g.11500-117C>G, g.66995-169insdelC	1063.612	1056.437
		g.11500-117C>G, g.32330-48A>G, g.66995-169insdelC	1071.769	1064.456
		g.8778G>A, g.11500-117C>G, g.32330-48A>G	1073.059	1068.879
		g.8778G>A, g.11500-117C>G, g.34425+102A>T	1075.583	1070.174
		g.4102+36T>G, g.8778G>A, g.34425+102A>T	1077.581	1072.194
근내지방도	2	g.4102+36T>G, g.11500-117C>G	15.0184	14.9371
		g.8778G>A, g.11500-117C>G	15.0224	15.0084
		g.4102+36T>G, g.32330-48A>G	15.0710	10.0247
		g.32330-48A>G, g.34425+102A>T	15.0710	15.0247
		g.32330-48A>G, g.66995-169insdelC	15.0710	15.0247
	3	g.4102+36T>G, g.11500-117C>G, g.32330-48A>G	14.8893	14.8357
		g.11500-117C>G, g.32330-48A>G, g.66995-169insdelC	14.8915	14.9337
		g.8778G>A, g.11500-117C>G, g.32330-48A>G	14.8980	14.9112
		g.4102+36T>G, g.8778G>A, g.11500-117C>G	14.9262	14.8823
		g.8778G>A, g.11500-117C>G, g.34425+102A>T	14.9559	14.9077

48A>G, g.34425+102A>T), (g.8778G>A, g.11500-117C>G, g.32330-48A>G)이 선별되었다.

3.4. SNPHarvester를 활용한 유전자 조합 선별 결과

한우의 주요 경제형질에 SNPHarvester를 적용하여 주요 유전자 조합을 찾아보았다. 이 때 사용된 스코어 함수는 값이고 P-값이 유의하고 동시에 값이 높은 상위 5개의 유전자 조합의 선별 결과를 표 3.3에 나타냈다.

SNPHarvester를 한우의 각 경제형질에 적용한 결과 2개 요인으로 결합된 SNP 그룹에서 종합적인 경제형질에 관한 유전자 조합으로 g.8778G>A과 g.11500-117C>G, g.11500-117C>G과 g.32330-48A>G을 선별하였다. 또한 3개 요인으로 결합된 SNP들 중에서 g,4102+38T>G, g.9889G>A, g.11500-117C>G가 종합적인 경제형질에 관계된 주요 유전자 조합으로 선별되었다. 따라서 SNPHarvester를 통해 한우의 종합적인 경제형질에 영향을 주는 우수 유전자 조합으로 (g.8778G>A, g.11500-117C>G), (g.11500-117C>G, g.32330-48A>G), (g,4102+38T>G, g.9889G>A, g.11500-117C>G)가

표 **3.3.** SNPHarvester를 활용한 한우의 각 경제형질에 대한 우수 유전자 조합 선별 결과

경제형질	요인수	유전자조합	χ^2
일당증체량	2	g.4102+36T>G, g.11500-117C>G	42.46
		g.8778G>A, g.11500-117C>G	36.30
		g.4102+36T>G, g.34425+102A>T	33.12
		g.11500-117C>G, g.32330-48A>G	29.74
		g.8778G>A, g.32330-48A>G	29.23
	3	g.4102+36T>G, g.11500-117C>G, g.34425+102A>T	52.36
		g.8778G>A, g.11500-117C>G, g.32330-48A>G	49.10
		g.4102+36T>G, g.8778G>A, g.11500-117C>G	48.72
		g.8778G>A, g.11500-117C>G, g.66995-169insdelC	44.54
		g.8778G>A, g.11500-117C>G, g.34425+102A>T	42.89
도체중량	2	g.4102+36T>G, g.8778G>A	56.15
		g.8778G>A, g.11500-117C>G	45.43
		g.8778G>A, g.66995-169insdelC	45.24
		g.8778G>A, g.34425+102A>T	44.31
		g.11500-117C>G, g.32330-48A>G	37.84
	3	g.8778G>A, g.34425+102A>T, g.66995-169insdelC	68.45
		g.4102+36T>G, g.8778G>A, g.66995-169insdelC	61.34
		g.4102+36T>G, g.8778G>A, g.11500-117C>G	59.52
		g.4102+36T>G, g.8778G>A, g.34425+102A>T	58.94
		g.32330-48A>G, g.34425+102A>T, g.66995-169insdelC	53.82
근내지방도	2	**g.8778G>A, g.11500-117C>G**	25.63
		g.8778G>A, g.32330-48A>G	25.06
		g.11500-117C>G, g.32330-48A>G	22.42
		g.11500-117C>G, g.34425+102A>T	18.82
		g.4102+36T>G, g.11500-117C>G	12.73
	3	g.8778G>A, g.11500-117C>G, g.32330-48A>G	44.99
		g.4102+36T>G, g.8778G>A, g.32330-48A>G	33.44
		g.8778G>A, g.32330-48A>G, g.34425+102A>T	28.47
		g.8778G>A, g.11500-117C>G, g.34425+102A>T	27.35
		g.4102+36T>G, g.8778G>A, g.11500-117C>G	27.26

표 **3.4.** E-MDR, D-MDR, SNPHarvester에서 최종 선별된 우수 유전자 조합 결과

	요인수	
	2개 결합	3개 결합
우수 유전자 조합	g.8778G>A, g.11500-117C>G	g.8778G>A, g.11500-117C>G, g.32330-48A>G

선별되었다.

상호작용 효과를 규명하기 위한 통계적 방법들을 통하여 한우의 경제형질에 관한 우수한 유전자 조합을 찾을 수 있었다. 그리고 E-MDR, D-MDR, SNPHarvester에서 각각 선별된 우수 유전자 조합들을 살펴본 결과 각 통계적 방법에서 거의 같은 유전자 조합들이 선별되었음을 볼 수 있었다. 표 3.4는 Lee (2009)의 연구에서 선택된 6개의 SNP들 중 이들 통계 방법에서 최종 선별된 우수 유전자 조합을 표로 나타낸 것이다.

표 3.4를 보면 2개 요인으로 결합된 SNP 그룹 중 E-MDR, D-MDR, SNPHarvester에서 공통으로 나온

표 **4.1.** CART 방법을 활용한 우수 유전자형 그룹 선별과 순열 검정 적용 및 결과

유전자 조합	경제형질	우수 유전자 형	평균(N)	t (p-value)	순열검정 (p-value)
g.8778G>A, g.11500-117C>G	일당증체량	**GGGC, GAGG, GGGG**	0.77(2225)	11.31 (0.0001)	< 0.001
	도체중량	**GGCC, GGGC, GAGG, GGGG**	329.99(2985)	11.08 (0.0001)	< 0.001
	근내지방도	AAGG, GAGC, **GGCC**	5.68(1595)	8.86 (0.0001)	< 0.001
g.8778G>A, g.11500-117C>G, g.32330-48A>G	일당증체량	GAGGAA, **GGGCAA, GAGGGA, GGGGAA, AAGGAA, GGCCGA**	0.76(2324)	11.78 (0.0001)	< 0.001
	도체중량	**GGGCAA, GAGGGA, GGGGAA, AAGGAA**	322.21(2229)	12.39 (0.0001)	< 0.001
	근내지방도	AAGG, GAGC, **GGCCAA** GGCCGA	6.10(1112)	10.22 (0.0001)	< 0.001

우수 유전자 조합은 (g.8778G>A, g.11500-117C>G)으로 선별되었고 3개 요인으로 결합된 SNP 그룹 중 E-MDR, D-MDR에서 공통으로 나온 우수 유전자 조합은 (g.8778G>A, g.11500-117C>G, g.32330-48A>G)으로 선별되었다. 3개 요인으로 결합된 유전자 조합은 세 방법에서 공통적으로 선별된 유전자 조합은 아니지만 3개 결합 중 가장 가능성 있는 유전자 조합으로 보고 선별된 유전자 조합에 대하여 우수 유전자형을 찾는다. 표 3.4에서 선별된 우수 유전자 조합에 대해 4장에서는 CART방법을 적용하여 한우의 경제형질의 가치를 높일 수 있는 우수 유전자형(genotype)을 선별하고 순열 검정을 활용하여 우수 유전자형의 통계적 유의성을 살펴본다.

4. 우수 유전자형 선별을 위한 CART 방법 적용 및 결과

각 통계적 방법들에서 선별된 우수 유전자 조합에서 우수 유전자형을 찾기 위하여 데이터 마이닝 기법 중 하나인 CART방법을 적용하여 평균이 높은 집단을 우수 유전자형으로 선별하였다. 또한 이렇게 선별된 유전자형이 통계적으로 유의한지 알아보기 위해 t 검정과 순열 검정을 실시한 것을 표 4.1에 나타냈다.

표 4.1을 살펴보면 2개 요인으로 결합된 (g.8778G>A, g.11500-117C<G) 유전자 조합 중 일당증체량에서 평균이 0.77인 GGGC, GAGG, GGGG가 선별되었다. 도체중량에서는 평균이 329.99인 GGCC, GGGC, GAGG, GGGG가 우수 유전자형으로 선별되었고 근내지방도에서는 평균이 5.68인 AAGG, GAGC, GGCC가 우수 유전자형으로 선별되었다. 2개 결합된 (g.8778G>A, g.11500-117C<G) 유전자 조합에서 최종적으로 GGGC, GAGG, GGGG, GGCC 유전자형이 공통적으로 나타났다. 이 유전자 조합에서 선별된 유전자형의 t 검정과 순열 검정의 p-value는 거의 0.0001로써 (g.8778G>A, g.11500-117C<G)에서 선별된 유전자형이 한우의 종합적인 경제형질에도 유의한 영향을 끼치는 것으로 나타났다. 또한 3개 요인으로 결합된 (g.8778G>A, g.11500-117C>G, g.32330-48A>G) 유전자 조합 중 일당증체량에서는 평균이 0.76으로 GGGCAA, GAGGGA, GAGGAA, GGGGAA, AAGGAA, GGCCGA가 우수 유전자형으로 선별되었고 도체중량에서는 평균이 322.21로 GGGCAA, GAGGGA, GGGGAA, AAGGAA가 선별되었다. 또한 근내지방도에서는 평균이 6.10으로 AAGGGA, GAGCGA, GGCCAA, GGCCGA가 우수 유전자형으로 선별되었다. 최종적으로 (g.8778G>A, g.11500-117C>G, g.32330-48A>G) 유전자 조합에서 GGGCAA, GAGGGA, GGGGAA, AAGGAA, GGCCGA 유전자형이 공통적으로 나타났다. 이 유전자형들의 일당증체량 역시 t 검정과 순열 검정의 p-value가 거

의 0.0001로써 한우의 종합적인 경제형질에 모두 유의한 영향을 주는 유전자형들임을 발견하였다. 따라서 한우의 경제형질의 가치를 높일 수 있는 우수한 유전자형은 유전자가 2개 요인으로 결합된 형태에서는 GGGC, GAGG, GGGG, GGCC, 3개 요인으로 결합된 형태에서는 GGGCAA, GAGGGA, GGGGAA, AAGGAA, GGCCGA으로 선별되었다.

5. 결론 및 토의

본 논문은 한우 4190두의 자료를 E-MDR, D-MDR, SNPHarvester 방법을 적용하여 6개의 SNP들 중 한우의 경제형질에 연관된 우수 유전자 조합을 선별하였다. E-MDR, D-MDR, SNPHarvester에서 찾은 우수 유전자 조합은 (g.8778G>A, g.11500-117C>G), (g.8778G>A, g.11500-117C>G, g.32330-48A>G)으로 선별하였다. 또한 우수 유전자 조합에서 우수 유전자형을 규명하기 위해 CART방법을 적용한 결과 한우의 종합적인 경제형질의 가치를 높일 수 있는 유전자형으로 우수 유전자 조합인 (g.8778G>A, g.11500-117C>G)에서 GGGC, GAGG, GGGG, GGCC를 우수 유전자형으로 선별하였고 (g.4102+36T>G, g.11500-117C>G, g.32330-48A>G)에서는 GGGCAA, GAGGGA, GGGGAA, AAGGAA, GGCCGA를 우수 유전자형으로 선별하였다. 이 유전자형이 한우의 경제형질에 유의한 영향을 주는지 알아보기 위해 t 검정과 순열 검정을 적용한 결과 우수 유전자 조합에서 선별된 우수 유전자형이 통계적으로 유의성을 가지는 것으로 밝혀졌다. 따라서 우리가 E-MDR, D-MDR, SNPHarvester에서 공통적으로 선별한 우수 유전자 조합에서 발견한 우수 유전자형이 한우의 경제형질에 유의한 영향을 끼친다는 것을 밝힐 수 있었다.

참고문헌

김태근 (2006). <u-Can 회귀분석>, 인간과 복지, 서울.

Barendse, W., Bunch, R., Thomas, M., Armitage, S., Baud, S. and Donaldson, N. (2004). The TG5 thyroglobulin gene test for a marbling quantitative trait loci evaluated in feedlot cattle, *Australian Journal of Experimental Agriculture*, **44**, 669–674.

Chung, Y. J., Lee, S. Y. and Park, T. S. (2005). Multifactor dimensionality reduction in the presence of missing observations, *Journal of Korea Statistical Society, Proceedings of the Autumn Conference*, **1**, 31–36.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Boostrap*, Champman & Hall/CRC.

Good, P. (2000). *Permutation Test: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer-Verlag, New York.

Kim, J. W, Park, S. I. and Yeo, J. S. (2003). Linkage mapping and QTL on chromosome 6 in Hanwoo(Korean Cattle), *Asian-Australasian Journal of Animal Sciences*, **16**, 1402–1405.

Lee, J. Y. and Lee, H. G. (2009). Multifactor Dimensionality Reduction(MDR) Analysis by Dummy Variables, *The Korean Journal of Applied Statistics*, **22**, 435–442.

Lee, J. Y., Lee, H. G. and Lee, Y. W. (2009). Expanded MDR algorithm based interaction effect discovery, *Journal of Applied Statistics*, **22**.

Lee, Y. S. (2009). Study on the identification of candidate genes and their haplotypes that are associated with growth and carcass traits in the QTL region of BTA6 in a Hanwoo population, Ph. D. Thesis, 1–94.

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

Snelling, W. M., Casas E., Stone, R. T., Keele, J. W., Harhay G. P., Benett G. L. and Smith T. P. L. (2005). Linkage mapping bovine EST-based SNP, *BMC Genomics*, **6**, 74–84.

Yang, C., He, Z., Wan, X., Yang, Q., Xue, H. and Yu, W. (2009). SNPHarvester: A filtering-based approach for detecting epistatic interactions in genome-wide association studies, *Bioinformatics*, **25**, 504–511.

Statistical Interaction for Major Gene Combinations

Jea-Young Lee[1] · Yong-won Lee[2] · Young-jin Choi[3]

[1]Department of statistics, Yeungnam University
[2]Department of statistics, Yeungnam University
[3]Department of statistics, Yeungnam University

(Received February 2010; accepted July 2010)

Abstract

Diseases of human or economical traits of cattles are occured by interaction of genes. We introduce expanded multifactor dimensionality reduction(E-MDR), dummy multifactor dimensionality reduction(D-MDR) and SNPHarvester which are developed to find interaction of genes. We will select interaction of outstanding gene combinations and select final best genotype groups.

Keywords: Expanded-MDR, Dummy-MDR, SNPHarvester, gene, genotype.

[1]Corresponding author: Professor, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea. Email : jlee@yu.ac.kr