Journal of the Korean Data & Information Science Society
2013, **24**(2), 277–287

http://dx.doi.org/10.7465/jkdi.2013.24.2.277
한국데이터정보과학회지

더미 다중인자 차원축소법에 의한 검증력과 주요 유전자 규명

여정수[1] · 라부미[2] · 이호근[3] · 이성원[4] · 이제영[5]

[12]영남대학교 생명공학부 · [345]영남대학교 통계학과

접수 2013년 1월 31일, 수정 2013년 2월 27일, 게재확정 2013년 3월 6일

요 약

광범위 유전자 관련 연구에서는 유전자-유전자 상호작용을 규명하는 것은 매우 중요하다. 최근 유전자-유전자 상호작용을 규명하는데에 대한 많은 연구가 진행되고 있다. 그 중 하나로 더미 다중인자 차원축소법이다. 이 연구의 목적은 모의실험을 통해 유전자-유전자 상호작용 파악하기 위한 더미 다중인자 차원축소의 검증력을 평가하는 것이다. 또한 이 방법을 적용하여 한우모집단에서 경제형질을 위한 단일 염기 다형성의 상호작용 효과를 확인하였다.

주요용어: 검증력, 경제형질, 더미-다중인자 차원축소.

1. 서론

사회 또는 교육, 생명공학 등 많은 분야에서 기술이 발전하고 복잡해짐에 따라 분석에 사용하고자 하는 변수의 수가 증가하고 그에 따라 상호작용을 고려해야하는 경우 또한 증가하였으며, 상호작용에 의한 영향이 중요시 되고 있다. 특히 광범위 유전자 관련 (genome-wide association; GWA)연구에서는 무수히 많은 유전자들을 이용하여 인간의 질병에 관련된 유전자를 찾아왔다. 이들을 분석하고 해석하기 위해 통계모형의 상호작용을 고려한 모형으로 선형모형과 같은 표준 통계 모형을 사용해왔다. 그러나 유전자의 상호작용처럼 변수가 많아질수록 상호작용의 조합이 많아지므로 모형이 복잡해지고, 종종 모수들의 상호작용에 대한 해석이 어려울 수 있다. 또한 Hosmer와 Lemeshow (2000)에 의하면 이런 유전자들을 가지고 모형화된 경우라도 많은 가능한 범주에 관측값이 없을 수도 있다. 그에 따라 이러한 문제들을 해결하기 위한 많은 방법들이 연구되어왔다. 그러한 연구들 중 인간 질병에 대한 상호작용을 결정하는 방법으로 Ritchie 등 (2001)과 Chung 등 (2005)이 다중인자 차원축소 (multifactor dimensionality reduction; MDR)방법을 제시하였고, Nelson 등 (2001)이 조합 분할 방법 (combinatorial partition method; CPM)을 제시하였으며, Culverhouse 등 (2004)이 제한적 분할 방법 (restricted partition method; RPM)을 제시하였다. 특히 다중인자 차원축소 방법은 상호작용에 대한 명확한 모형의 가정이 없는 비모수적인 방법으로 적당한 상위-하위 차수의 데이터로 복잡한 관계를 밝힐 수 있었다. MDR방법은 사례항목과 대조항목의 비율을 통해 독립변수의 범주를 '상위'집단과 '하위'집단으로 분류한 후 목표변수에 대한 오분류율을 비교하여 분석한다. 그러나 이 MDR방법은 사례-대조로 이분화 된 데이터에 대해 사용하는 방법으로 연속형 데이터에는 적용할 수 없다. 이러한 문제점을 해결

[1] (712-749) 경북 경산시 대학로 280, 영남대학교 생명공학부, 교수.
[2] (712-749) 경북 경산시 대학로 280, 영남대학교 생명공학부, 대학원생.
[3] (712-749) 경북 경산시 대학로 280, 영남대학교 통계학과, 대학원생.
[4] (712-749) 경북 경산시 대학로 280, 영남대학교 통계학과, 강사.
[5] 교신저자: (712-749) 경북 경산시 대학로 280, 영남대학교 통계학과, 교수. E-mail: jlee@yu.ac.kr

하기 위한 방법으로 Lee 등 (2009a)이 더미변수를 활용한 다중인자 차원축소 (D-MDR)방법을 제시하였다. D-MDR방법은 더미변수를 활용한 회귀분석방법을 응용하여 기존 MDR방법에 적용하는 방법으로 각 상호작용 조합을 더미변수로 하는 회귀모형에서 회귀계수의 해석상 의미를 이용하여 각 범주들 사이의 상호간의 차이를 비교하여 범주를 '상위'집단과 '하위'집단으로 분류하는 방법을 기존 MDR방법에 적용하는 형태이다. 본 연구에서는 앞서 제시된 D-MDR 방법의 검정력을 검증하기 위해 모의실험을 통한 검정력 (power)을 평가하였다 (Choi, 2012). 모의실험은 Ritchie 등 (2003)이 MDR방법의 검정력에 대해 연구한 방법과 Culverhouse 등 (2004)이 RPM방법의 검정력에 대해 연구한 방법을 참고로 하여 epiatatic 모형을 본 모의실험의 기본 모형으로 사용하였으며, 여러 데이터 특성에서 D-MDR방법의 검정력을 확인하기 위해 데이터의 분포와 개체수를 달리 하여 실험하였다. 또한 검정력의 측도로는 William 등 (2008)등이 MDR방법의 평가 측도로 사용한 "Detection"을 사용하였으며, 이는 해당 선별 방법이 모든 상호작용 조합에서 정확하게 선별하는 능력을 뜻하는 것으로 본 연구에서는 10개의 SNPs의 조합인 45개 조합 중 상호작용 효과로 정의한 1개의 조합을 정확하게 선별 (detection)하는 정확도 (accuracy)를 말한다. "Detection"에 해당하는 정확도의 평가를 통해 D-MDR방법의 특성과 선별된 상호작용 조합의 신뢰성을 확인하였다. 아울러 평가된 D-MDR방법을 한우의 경제형질자료에 적용하여 육량과 육질을 나타내는 형질인 일당증체량 (average daily gain; ADG)과 근내지방도 (marbling score; MS)에 관련하는 우수 유전자 조합을 선별하였다 (Lee 등, 2011).

본 연구는 연구목적인 D-MDR방법의 모의실험을 통한 검정력 평가를 위해 먼저 2절에서 상호작용을 선별하는 방법인 D-MDR방법에 대해 소개한 후 3절에서 모의실험을 통해 D-MDR방법의 검정력으로 정확도를 평가한다. 4절에서는 3절에서 검증한 D-MDR방법에 한우의 경제형질 데이터를 적용하여 우수 유전자 조합을 선별하고, 5절에서 본 연구의 결론 및 개선점에 대해 나타내었다.

2. 더미변수를 활용한 D-MDR방법의 소개

MDR (multifactor dimensionality reduction) 방법은 일반화된 선형 모형인 전통적인 통계 기법과는 달리 어떤 모수에 대한 추정과 genetic 모형의 가정을 요구하지 않는다 (다시 말하면 특별한 유전형질 모형에 대한 가정이 필요 없다). Ritchie 등 (2003)에 의하면 MDR 모형은 사례와 대조의 이분화된 데이터에 적합 가능하다. 따라서 연속형 자료에 적용 할 수 없다. 이러한 문제를 해결하기 위해 제안된 방법인 더미변수를 활용한 D-MDR방법을 소개한다. 더미변수를 활용한 D-MDR방법은 사례-대조 데이터에만 적용 가능한 MDR방법의 문제점을 해결하기 위해 제안된 방법으로 더미변수를 활용한 회귀분석에서 회귀계수의 해석상 의미를 활용하여 이분화 시키는 방법을 적용한 방법이다. Figure 2.1은 연속형 데이터에 대한 D-MDR방법 (Lee 등, 2009a)의 분석과정을 나타낸다.

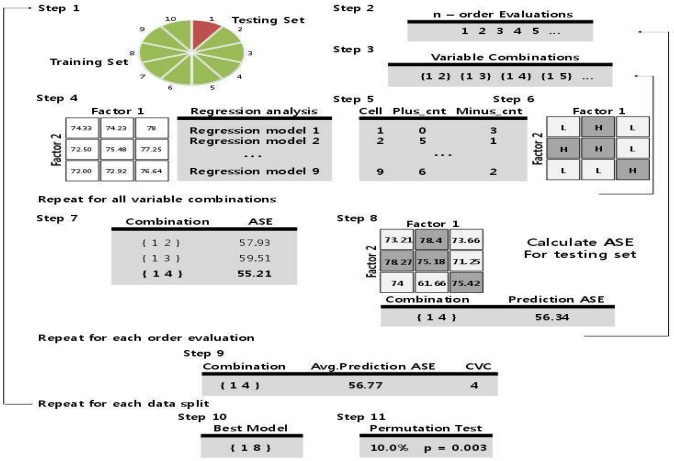

Figure 2.1 The general steps using a dummy variable method to implementing the MDR method.

이때, 선별된 우수 SNPs 조합에 대한 통계적 유의성을 검정하기 위해 순열 검정 (permutation test; Good, 2000)을 실시하였으며, 그 과정은 Figure 2.1과 같다. 2절에서는 이분형 자료에서 우수 상호작용 효과를 선별하는 D-MDR방법의 절차를 소개하였다.

3. 모의실험을 통한 D-MDR방법의 정확성 검정

3.1. 모의실험 자료

목표변수가 연속형 자료인 경우 상호작용을 분석하기 위해 제안된 D-MDR 방법을 모의실험을 통해 평가함에 있어 실제 분석자료인 유전자와 연속형 형질자료에 적합한 실험을 수행하기위해 다음과 같이 모의실험 자료를 생성하였다. 2가지의 다른 2형질 상호작용 모형 (체크보드 모형, 대각화 모형)을 사용한 이분형 데이터를 연속형 데이터 생성에 활용하였다. epistatic 모형을 사용한 것은 실험의 결과로 나타난 효과가 주효과에 의한 결과인지 상호작용에 의한 결과인지 또는 모두에 의한 결과인지 평가하기 어렵기 때문이다 (Ritchie 등, 2003). 또한 여러 연구에서 사용된 epistatic 모형들 중에서 본 연구와 같이 연속형 자료를 분석하는 방법인 RPM방법의 모의실험 (Culverhouse 등, 2004)에서 사용된 체크보드 (checkerboard) 모형과 대각화 (diagonal) 모형을 사용하였다. Figure 3.1과 Table 3.1, Table 3.2는 각 모형의 형태와 모형특성을 나타낸다.

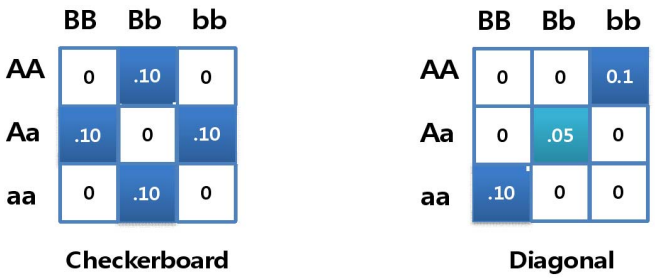

Figure 3.1 Two classes of purely epistatic models used for simulations ($p = 0.5$, $q = 0.5$)

Table 3.1 Checkerboard model combinations and the individual in each case the ratio of the expression

	Case expression ratio in combination			Individual
	BB (.25)	Bb (.50)	bb (.25)	expression ratio
AA (.25)	0.0	1.0	0.0	0.5
Aa (.50)	1.0	0.0	1.0	0.5
aa (.25)	0.0	1.0	0.0	0.5
Individual expression ratio	0.5	0.5	0.5	

Table 3.2 Diagonal model combinations and the individual in each case the ratio of the expression

	Case expression ratio in combination			Individual
	BB (.25)	Bb (.50)	bb (.25)	expression ratio
AA (.25)	0.00	0.00	1.00	0.25
Aa (.50)	0.00	0.50	0.00	0.25
aa (.25)	1.00	0.00	0.00	0.25
Individual expression ratio	0.25	0.25	0.25	

Table 3.1과 Table 3.2에서 조합에 의한 사례의 발현 비율은 다르지만 각각 개별적인 형태에서 사례의 발현 비율은 동일함을 알 수 있다. 이는 조합에 의한 영향은 존재하지만 개별적인 영향은 없다는 것을 의미한다. 이러한 특성과 Hardy-Weinberg 평형을 유지하는 모형을 epistatic 모형이라 한다 (Ritchie 등, 2003).

다음으로 본 연구는 연속형 자료에 대한 모의실험으로 각 그룹에 따른 연속형 자료의 분포에 따라 모형의 정확도가 달라질 수 있으므로 총 4가지의 분포를 가지고 실험하였다. 각각의 분포는 Culverhouse 등 (2004)이 실험한 RPM의 모의실험에서 사용된 분포 중 4개의 분포를 활용하였으며, 형태는 유지하면서 평균과 표준편차의 경우 실제 적용에 사용될 데이터의 특성을 반영하여 생성하였다. Table 3.3과 Figure 3.2은 각 분포의 특성과 형태를 나타낸다.

Table 3.3 RPM and D-MDR characteristics of the distribution used in simulation

	RPM simulation		D-MDR simulation		
	Case group	Control group	Case group	Control group	Region (%)
mean	0	1	312.77	383.93	
SD1	0.25	0.25	17.79	17.79	4.56
SD2	0.50	0.50	35.58	35.58	31.80
SD3	1.00	1.00	71.16	71.16	61.80
SD4	2.00	2.00	142.32	142.32	80.20

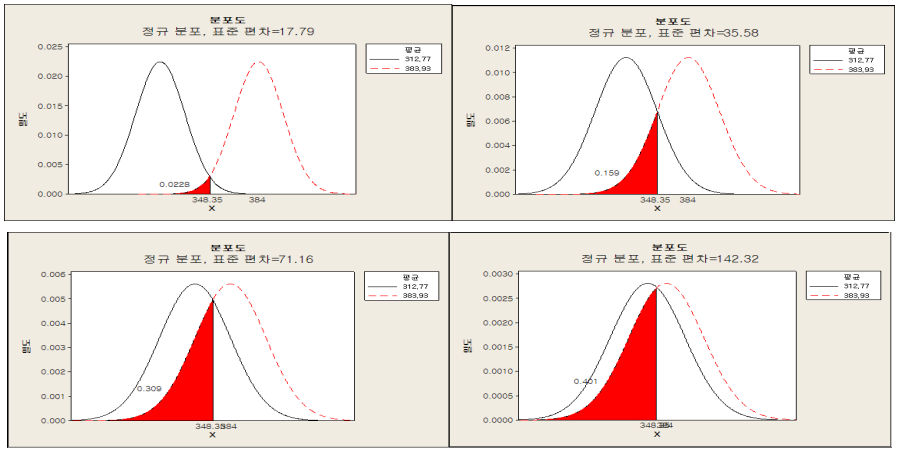

Figure 3.2 Simulated distributions used in the form of each normal group

모형의 정확도는 개체수에 따라 차이가 날 수 있으므로 개체수를 3가지의 형태 (N = 400, 1000, 2000)로 데이터를 생성하여 실험하였다. 또한 각 실험마다 10개의 SNPs를 사용하여 총 45개의 SNPs 조합 (2개의 조합)을 비교하여 두 모형에 의해 정의된 조합이 얼마나 정확하게 선별되는지를 실험하였다. 각 실험은 10번의 반복을 통해 이루어졌으며, 그룹의 빈도는 균형화되도록 생성되었다. 즉, 각 방법에서 모형 (2가지), 표준편차 (4가지), 데이터수 (3가지), 반복수 (10회)를 가지고 총 240개의 데이터 셋 (2400개의 교차타당성 데이터 셋)을 생성하여 실험하였다. 다음의 절차와 Figure 3.3은 SAS ver. 9.13 프로그램을 사용한 데이터 생성 과정을 나타낸다.

Figure 3.3의 과정에서 생성된 각 모형의 데이터의 일부 형태는 Figure 3.4와 같이 나타났으며, 생성된 데이터를 모의실험에 사용하였다. 3.2절에서는 모의실험을 통해 알고자 하는 모형의 정확도에 대한 측도를 나타낸다.

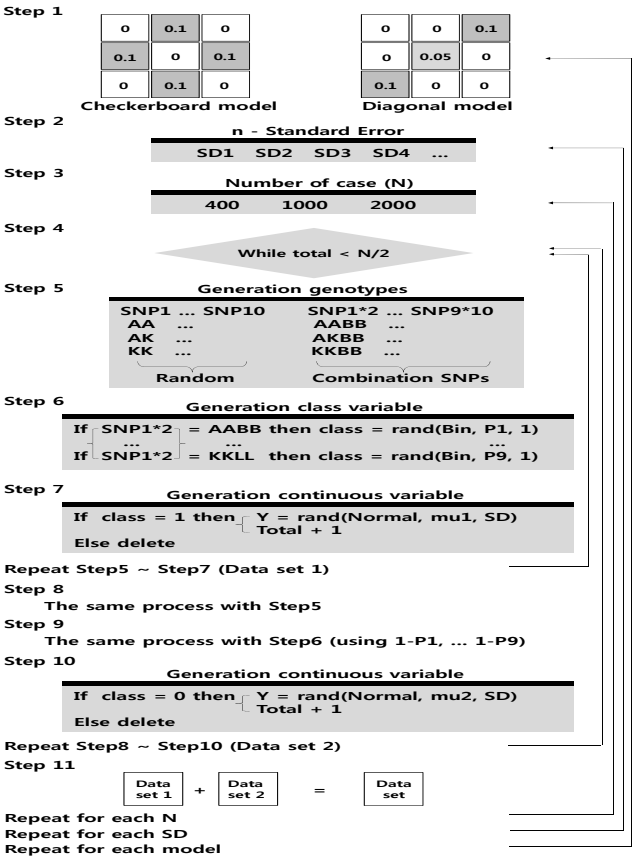

Figure 3.3 Full data generation algorithm for the D-MDR simulation (SAS ver. 9.13 used)

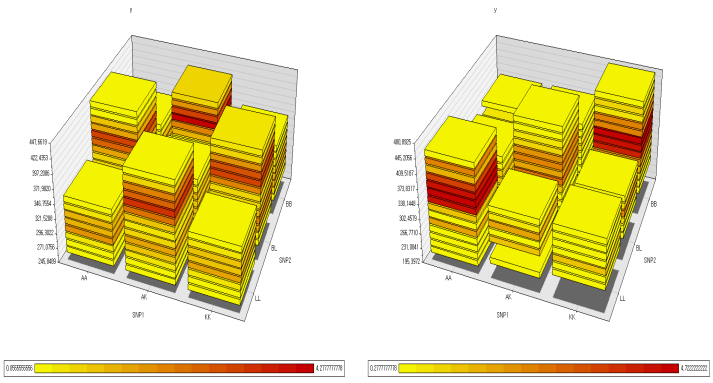

Figure 3.4 Data distribution generated by Epistatic model (n=2000, SD1).

3.2. 평가 측도

D-MDR방법의 검정력을 평가하기 위한 측도로 William 등 (2008)이 MDR방법의 평가 측도로 사용한 “Detection”을 사용하였으며, 이는 해당 선별 방법이 모든 상호작용 조합에서 정확하게 선별하는 능

력을 나타내는 것이다. 본 연구에서는 10개의 SNPs의 조합인 45개 조합 중 상호작용 효과로 정의한 1개의 조합을 정확하게 선별하는 정확도 (accuracy)를 말한다. 즉, 각 데이터 셋에서 생성된 그룹과 연속형 자료는 모두 SNP1과 SNP2의 조합인 SNP1*2에 의해서 epistatic 모형을 기준으로 생성된 것으로 이 외의 조합은 연속형 자료와 무관하게 무작위 유전자형을 가진다. 따라서 D-MDR 방법에 각 데이터 셋을 적용하여 분석한 결과 SNP1*2가 우수한 상호작용을 가진 조합으로 선택된다면 올바른 선택으로 볼 수 있다. 이러한 평가측도인 정확도를 위해 2개의 epistatic 모형, 4개의 표준편차, 3개의 개체수에 따른 총 24가지 상황별로 10번의 반복에서 정확도를 비교하였다. 3.3절에서는 3.1절에서 생성한 모의실험자료를 D-MDR방법에 적용한 결과를 위에서 나타낸 "Detection"에 해당하는 정확도를 통해 평가하고 결과를 통해 나타나는 특성을 나타낸다.

3.3. 모의실험 결과

모의실험의 결과는 3.2절에서 설명한 검정력의 측도인 "Detection"에 해당하는 값을 얻기 위해 D-MDR방법의 절차인 교차타당성 시행을 통해 하나의 결과를 얻고, 그 과정을 10번 반복하여 정확도를 계산하였다. Table 3.4는 24가지의 실험 중 3가지의 결과에 해당하는 것으로, 체크보드 모형의 표준편차 SD3에서 각 개체수 (N=400, 1000, 2000)에서 10회 반복한 결과이다. 또한 Table 3.6의 결과 각 개체수 집단에서 10회 반복을 통해 우수 유전자조합이 선택되었으며 정확도는 각각 80%, 100%, 100%로 나타났다. 각각의 경우에서 CVC (cross validation consistency)값을 보면 평균 CVC 값이 2.30, 5.00, 8.90으로 개체수가 많아질수록 정확도와 CVC값이 좋아짐을 알 수 있다. Table 3.5과 Table 3.6에서 각 epistatic 모형에서 모의실험 결과를 종합하여 나타낸다. 모의실험 결과 Table 3.5와 같이 모형의 정확도를 확인하였다. 두 epistatic 모형에서의 결과는 큰 차이는 없었으나 대각화 모형에서의 정확도가 조금 더 높았으며, 두 그룹의 표준편차가 작을수록 CVC값과 정확도가 증가하였다. 또한 개체수가 증가할수록 CVC값과 정확도가 증가하였다. 이는 MDR방법이 교차타당성을 통해 정확도의 향상을 유도하지만 그럴 경우 테스팅 셋이 전체 데이터의 1/10 크기로 줄어들게 되므로 개체수가 적은 경우 불안정한 결과를 가져오는 것으로 판단된다.

Table 3.4 Accuracy results by Iterations for each population in checkerboard model (SD3)

Standard Divation	N	repitation	Superior gene combinations				ACC (%)
			Avg. ASE	Avg. P_ASE	CVC	selected gene combination	
SD3	400	1	5230.18	4927.29	4.00	1_2	80
		2	5919.72	5549.78	2.00	9_10	
				...			
		9	5290.60	5108.96	2.00	1_2	
		10	5905.45	5659.30	1.00	1_2	
		mean	5454.51	5160.13	2.30		
	1000	1	6229.49	6140.71	5.00	1_2	100
		2	6095.00	6032.41	4.00	1_2	
				...			
		9	5774.19	5635.70	5.00	1_2	
		10	5492.33	5393.78	1.00	1_2	
		mean	5909.91	5819.58	5.00		
	2000	1	5785.33	5727.64	8.00	1_2	100
		2	5727.64	5819.24	10.00	1_2	
				...			
		9	5814.35	5782.13	8.00	1_2	
		10	6125.74	6064.65	8.00	1_2	
		mean	5891.77	5855.71	8.90		

Table 3.5 Accuracy results for checkerboard model simulations

Model Types	N	D - MDR results			ACC (%)
		Avg. ASE	Avg. P_ASE	Avg. CVC	
Checkerborad					
	400	1068.13	1033.29	8.9	100
SD1	1000	1052.76	1039.81	10.0	100
	2000	1074.14	1064.46	10.0	100
	400	2053.85	1978.20	5.8	100
SD2	1000	2006.53	1971.66	9.8	100
	2000	2039.29	2023.77	10.0	100
	400	5454.51	5160.13	2.3	80
SD3	1000	5909.91	5819.58	5.0	100
	2000	5891.77	5855.71	8.9	100
	400	2053.85	1978.20	5.8	100
SD2	1000	2006.53	1971.66	9.8	100
	2000	2039.29	2023.77	10.0	100

Table 3.6 Accuracy results for diagonal model simulations

Model Types	N	D - MDR results			ACC (%)
		Avg. ASE	Avg. P_ASE	Avg. CVC	
Diagonal					
	400	965.79	933.68	9.7	100
SD1	1000	922.38	903.04	10.0	100
	2000	938.59	929.52	10.0	100
	400	1916.55	1842.29	6.5	100
SD2	1000	1853.33	1825.07	9.9	100
	2000	1823.27	1806.22	10.0	100
	400	5454.51	5160.13	2.3	80
SD3	1000	5909.91	5819.58	5.0	100
	2000	5891.77	5855.71	8.9	100
	400	20347.55	19166.88	1.6	40
SD2	1000	21003.07	20578.06	1.9	80
	2000	20760.21	20491.69	2.8	100

4절에서는 본 연구에서 검증한 D-MDR방법을 실제 한우의 경제형질 자료에 적용하여 한우 경제형질에 우수한 상호작용 효과를 지닌 유전자 조합을 선별하였으며, 분석 과정에서 3장의 모의실험결과를 토대로 개체수의 보완을 위해 붓스트랩 (Efron 등, 1993; Sohn 등, 2012) 방법을 사용해 개체수를 보완하여 우수한 상호작용 효과를 지닌 유전자 조합을 선별하였다.

4. 한우의 경제형질에서 우수 유전자 조합 선별

4.1. 실험 자료

Lee 등 (2009)에 의해 소개된 한우의 수많은 DNA 마커들 중 경제형질을 조절하는 후보 DNA 마커로 연구에 사용된 한우는 농협중앙회 한우개량사업소의 후대검정집단인 30차에서 35차 국가 후대검정우 집단 476두로 구성되어졌다. 국가 후대검정우 집단의 개체들은 국가 씨수소 선발검정에서 당대검정으로 선발된 50두의 후보씨수소를 한우 암소개량농가에서 교배하여 생산된 수송아지이며, 다음과 같이 경제형질 연관 후보 DNA 마커를 선정하였다. 한우 6번 염색체에서 Kim 등 (2003)이 보고한 3개의 QTL 중 BMS1242와 ILST035는 일당증체량, 등지방두께, 등심단면적과 근대지방도에서 LOD (logarithm of odds)값이 3.0이상으로 나타났으며, BM4311은 근내지방도에서 LOD값이 3.0이상으로

나타났다 (Kim 등, 2003). 따라서 경제형질에 연관되는 유전자로 판단되는 QTL 양쪽으로 10cM 정도의 microsatellite를 선정하여 영역으로 지정하면, 육질과 육량에 대한 QTL과 연관된 유전자가 존재할 것으로 판단하였다. 이러한 2개의 QTL 영역에서 EST_based SNP 연관지도 (Snelling 등, 2005)에서 총 33개의 SNPs를 확인하였다. 발굴된 33개 SNPs 중 후보 유전자로 판단되어지는 LOC534614 유전자내의 SNPs 20개를 발견하였고, 이 중 대립유전자의 빈도가 0.1 미만이거나 유전자형의 빈도가 치우친 SNP인 g.934425+29T, g.34425+19T>C, g.-8606+137C>T를 제외하고 분석하였다.

이들을 제외한 17개의 SNPs들 간에 강한 연관불평형을 구성하고 있다는 것은 유전자가 하나의 변이로 작용하는 것이 아니라 여러개의 변이가 서로 조합되어 작용함을 의미한다. 또한 17개의 SNPs들의 상호작용을 모두 분석하는 것보다 htSNP (haplotype-tagging SNP)을 사용하여 분석하면 적은 수의 분석으로도 영향력을 확인 할 수 있음을 나타내었으며, 그에 따라 Table 4.1과 같이 6개의 SNPs (g.4102+36T>G, g.8778G>A, g.11500-117C>G, g.32330-48A>G, g.34425+102A>T, g.66995-169insdelC)를 확인 하였다. 이렇게 발견된 6개의 SNPs들을 이용하여 E-MDR방법과 D-MDR방법에 적용한 결과를 4.2절에 나타내었다.

4.2. 적용 결과

한우의 경제형질 육량과 육질에 해당하는 일당증체량 (average daily gain; ADG)과 근내지방도 (marbling score)에 대해 D-MDR에 적용한 결과를 각각 Table 4.1과 Table 4.2, Table 4.3에 나타내었다. D-MDR방법에 적용한 결과 개별적인 SNP의 효과와 2개의 SNP조합에 의한 효과는 각 경제형질에서 g.34425+102A>T와 g.8778G>A, g.11500-117C>G로 일치하였으나 3개의 SNP조합에 의한 결과에서는 일당증체량의 경우 g.8778G>A, g.11500-117C>G, g.34425+102A>T의 조합, 근내지방도의 경우 g.8778G>A, g.11500-117C>G, g.66995-169insdelC의 조합으로 서로 다르게 나타났다. 순열 검정의 결과에서는 근내지방도에 대한 개별적인 효과를 제외하고는 모두 유의하게 나타났다.

Table 4.1 Average daily gain and marbling score of a single SNP on the effect of D-MDR method

SNP combination	average daily gain (ADG)		marbling score (MS)	
	ASE	P_ASE	ASE	P_ASE
g.8778G>A	0.007782	0.007748	15.2783	15.2491
g.32330-48A>G	0.007781	0.007753	15.3119	15.2322
g.11500-117C>G	0.007811	0.007772	15.2038	15.1495
g.34425+102A>T	0.007778	0.007740	15.0710	15.0247
g.4102+36T>G	0.007820	0.007776		
g.66995-169insdelC	0.007832	0.007751		
p-value		0.147600		0.0097

Table 4.2 2 SNPs combination of average daily gain and marbling score on the effect of D-MDR method

SNP combination	average daily gain (ADG)		marbling score (MS)	
	ASE	P_ASE	ASE	P_ASE
g.8778G>A, g.32330-48A>G	0.007989	0.007927	15.2554	15.2189
g.8778G>A, g.11500-117C>G	0.007821	0.007775	15.0184	14.9371
g.8778G>A, g.34425+102A>T	0.007978	0.007917	15.0710	15.0247
g.8778G>A, g.4102+36T>G	0.008003	0.007954	15.2751	15.2281
g.8778G>A, g.66995-169insdelC	0.007988	0.007945	15.2014	15.1659
	...			
g.34425+102A>T, g.66995-169insdelC	0.007986	0.007929	15.0710	15.0247
g.4102+36T>G, g.66995-169insdelC	0.008042	0.007979	15.2135	15.1802
p-value		0.010770		0.0150

Table 4.3 3 SNPs combination of average daily gain and marbling score on the effect of D-MDR method

SNP combination	average daily gain (ADG)		marbling score (MS)	
	ASE	P_ASE	ASE	P_ASE
g.8778G>A, g.32330-48A>G, g.11500-117C>G	0.007580	0.007542	14.9262	14.8823
g.8778G>A, g.32330-48A>G, g.34425+102A>T	0.007720	0.007699	15.0398	14.9804
g.8778G>A, g.32330-48A>G, g.4102+36T>G	0.007718	0.007674	15.2710	15.2440
g.8778G>A, g.32330-48A>G, g.66995-169insdelC	0.007713	0.007686	15.1547	15.0992
g.8778G>A, g.11500-117C>G, g.34425+102A>T	0.007584	0.007535	14.8893	14.8357
g.8778G>A, g.11500-117C>G, g.4102+36T>G	0.007540	0.007504	15.0086	14.9177
g.8778G>A, g.11500-117C>G, g.66995-169insdelC	0.007524	0.007479	15.0505	15.0003
...				
g.11500-117C>G, g.4102+36T>G, g.66995-169insdelC	0.007653	0.007620	15.0621	15.0375
g.34425+102A>T, g.4102+36T>G, g.66995-169insdelC	0.007663	0.007618	15.0852	15.0562
p-value		0.020580		0.0153

D-MDR방법에 의한 결과를 가장 우수한 SNPs 조합과 순열 검정결과를 통해 유의한 SNPs 조합으로 정리하면 Table 4.4와 같이 나타났다.

Table 4.4 Final major SNPs selected by each method, a combination

Economic Traits		Using D-MDR method selected the best combination of SNP
ADG	2 combination	g.8778G>A, g.11500-117C>G
	3 combination	g.8778G>A, g.11500-117C>G, g.66995-169insdelC
MS	2 combination	g.8778G>A, g.11500-117C>G
	3 combination	g.8778G>A, g.11500-117C>G, g.34425+102A>T

각 조합에서 g.8778G>A, g.11500-117C>G 조합이 공통으로 포함되어 있음을 확인하였다. 따라서 각 방법에서 우수 SNPs 조합으로 선별된 g.8778G>A, g.11500-117C>G 조합이 한우의 육량 (일당증체량)과 육질 (근내지방도)을 가치를 높일 수 있는 우수 유전자 조합으로 나타났다.

5. 결론 및 토의

우리는 광범위 유전자 관련 (genome-wide association; GWA)연구에서 많은 유전자들을 이용하여 인간의 질병에 관련된 상호결합 유전자를 찾는 방법으로 제시된 MDR방법과 이분형 자료에만 적용 가능한 MDR방법의 한계를 극복하기 위한 방법으로 제시된 D-MDR방법을 소개하고, D-MDR 방법에 의한 우수 유전자 조합 선별결과에 대해 신뢰성 등을 확인하기 위해 모의실험을 통해 검정력에 대한 평가를 하였다. 선별능력 (detection)에 해당하는 정확도 (accuracy)를 통해 평가하였으며, 모의실험 결과 평균 정확도 92.08%로 높게 나타났다. 특히 개체수가 충분히 크고 (n >1000), 두 그룹의 분포에서 겹치는 면적이 약60% 미만일 경우 거의 대부분의 경우 정확하게 조합을 선택하였다. 즉, 모의실험 결과 우수 상호작용 조합을 선별하는 방법 D-MDR방법의 신뢰성을 확인하였다. 또한 모의실험을 통해 검정된 D-MDR방법을 적용하여, Lee 등 (2009)이 haplotype 분석을 통해 선별한 한우의 경제형질에 우수한 6개 SNPs로 한우의 육량과 육질에 해당하는 일당증체량과 근내지방도에 영향을 주는 우수한 SNPs 조합을 선별하였다. 그 결과 한우의 종합적인 (육량+육질) 경제형질에 관련된 우수한 유전자 조합은 g.87 78G>A, g.11500-117C>G의 조합으로 선별되었다. 즉, 이런 유전자 조합을 추가적으로 분석을 통해 유전자형을 개량한다면 더욱 가치가 높은 한우를 개발할 수 있을 것으로 판단된다. 하지만 MDR 방법의 한계점인 SNP의 수가 너무 많은 조합의 경우 계산이 복잡하고 과다비용등의 문제점도 있다. 추후, D-MDR방법은 우수한 상호작용 조합만을 선별하는 방법으로 우수한 유전자형 (genotypes)을 선별하기 위해서는 의사결정나무 (decision tree)방법 등의 추가적인 분석을 하여야한다. 따라서 D-MDR방법과 다른 방법들을 결합하는 하이브리드 (hybrid)방법등에 대한 연구가 필요할 것으로 사료된다.

참고문헌

Bush, W. S., Edwards, T. L., Duek, S. M., McKinney, B. A. and Ritchie, M. D. (2008). Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *Bio Medical Central Bioinformatics*, **16**, 238.

Choi, Y. H. (2012). Power study for 4 x 4 graeco-latin square design. *Journal of the Korean Data & Information Science Society*, **23**, 683-691.

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

Culverhouse, R., Klein, T. and Shannon, W. (2004). Detecting epistatic interactions contributing to quantitative traits. *Genetic Epidemiology*, **27**, 141-152.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*, Chapman & Hall/CRC, New York.

Good, P. (2000). *Permutation test: A ractical guide to resampling method for testing hypotheses*, Springer-Verlag Berlin and Heidelberg GmbH & Co, New York,

Hosmer, D. W. and Lemeshow, S. (2000.) *Applied logistic regression*, John Wiley & Sons, 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.

Kim, N. J. and Choi, K. H. (2012). Lipid metabolic effects of caffeine using meta-analysis. *Journal of the Korean Data & Information Science Society*, **23**, 649-656.

Lee, J. H. and Yeo, J. S. (2011). Estimation of genetic parameters using real-time ultrasound measurements in Hanwoo. *Journal of the Korean Data & Information Science Society*, **22**, 1145-1152.

Lee, J. Y. and Lee, H. G. (2009a). Multifactor dimensionality reduction (MDR) analysis by dummy variables. *The Korean Journal of Applied Statistics*, **22**, 435-442.

Lee, J. Y. and Lee, H. G. (2009b). A study on the comparison between E-MDR and D-MDR in continuous data. *Communications of the Korean Statistical Society*, **16**, 579-586.

Lee, J. Y., Lee, H. G. and Lee, Y. W. (2009). Study gene interaction effect based on expanded multifactor dimensionality reduction algorithm data. *The Korean 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, Yeungnam University, Kyungbuk.

Lee, Y. S., Bae, J. H., Lee, J. Y., Park, H. S. and Yeo J. S. (2008). Identification of candidate SNP for economic traits on chromosome 6 in Korean cattle. *Asian-Australasian Journal of Animal Sciences*, **21**, 1703-1709.

Nelson, M. R., Kardia, S. L., Ferrell R. E. and Sing, C. F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, **11**, 458-470.

Ritchie, M. D., Hahn, L. W. and Moore, J. H. (2003). Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phencopy, and Genetic Heterogeneity. *Genetic Epidemiology*, **24**, 150-157.

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

Journal of the Korean Data & Information Science Society
2013, **24**(2), 277–287

http://dx.doi.org/10.7465/jkdi.2013.24.2.277
한국데이터정보과학회지

Power and major gene-gene identification of dummy multifactor dimensionality reduction algorithm

Jungsou Yeo[1] · Boomi La[2] · Ho-Guen Lee[3] · Seong-Won Lee[4] · Jea-Young Lee[5]

[12]School of Biotechnology, Yeungnam University
[345]Department of statistics, Yeungnam University

Received 31 January 2013, revised 27 February 2013, accepted 6 March 2013

Abstract

It is important to detect the gene-gene interaction in GWAS (genome-wide association study). There have been many studies on detecting gene-gene interaction. The one is D-MDR (dummy multifoactor dimensionality reduction) method. The goal of this study is to evaluate the power of D-MDR for identifying gene-gene interaction by simulation. Also we applied the method on the identify interaction effects of single nucleotide polymorphisms (SNPs) responsible for economic traits in a Korean cattle population (real data).

Keywords: D-MDR, economic traits, power.

[1] Professor, School of Biotechnology, Yeungnam University, Kyungsan 712-749, Korea.
[2] Graduate student, School of Biotechnology, Yeungnam University, Kyungsan 712-749, Korea.
[3] Graduate student, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea.
[4] Instructor, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea.
[5] Corresponding author: Professor, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea. E-mail: jlee@yu.ac.kr