응용통계연구 (2009)
22(5), 1085–1096

환자-대조군 연구에서 인구집단 층화가 일배체형 경향성 검정에 미치는 영향

김진흠[1] · 강대룡[2] · 임현선[3] · 남정모[4]

[1]수원대학교 통계정보학과, [2]연세대학교 세브란스병원 임상시험센터
[3]연세대학교 의과대학 의학통계학과, [4]연세대학교 의과대학 예방의학교실

(2009년 8월 접수, 2009년 9월 채택)

요 약

환자-대조군 연관성 연구에서 후보 유전자와 질병이 연관되어 있지 않더라도 인구집단 층화로 인해 가짜 연관성이 발생할 수도 있다. 본 연구에서는 일배체형에 기초한 환자-대조군 연관성 연구에서 인구집단 층화로 인한 가짜 연관성을 해결하기 위한 방법으로, Zaykin 등 (2002)이 제안한 일배체형 경향성 모형에 인구집단 층화에 대한 정보를 추가하고자 한다. Zaykin 등 (2002)의 모형과 제안한 모형에 기초한 일배체형의 유의성 검정에서 인구집단 층화와 인구집단에 대한 관측 오차가 제1종 오류율에 미치는 영향을 모의실험을 통해 살펴보았다. 인구집단이 층화되어 있지만 각 개체가 속한 인구집단을 정확히 알 수 있을 때, Zaykin 등 (2002)의 모형에 기초한 검정은 제1종 오류율을 잘 조절하지 못했지만 본 연구에서 제안한 모형에 기초한 검정은 제1종 오류율을 잘 조절하는 것으로 나타났다. 그러나 인구집단이 층화되어 있고 관측 오차가 존재하면 제안한 모형에 기초한 검정도 제1종 오류율을 조절하지 못하고 명목 유의수준보다 큰 값을 갖는 것으로 나타났다. 따라서 단일염기다형성에 기초한 환자-대조군 연관성 연구와 마찬가지로 일배체형에 기초한 환자-대조군 연관성 연구에서도 인구집단 층화에 대한 정보를 갖고 있다할지라도 그 속에 관측 오차가 존재하면 위양성을 피하기 어렵다는 것을 알 수 있었다.

주요용어: 인구집단 층화, 가짜 연관성, 위양성, 일배체형 경향성 검정, 관측 오차.

1. 서론

유전역학(genetic epidemiology) 분야에서 질병과 관련된 유전자(disease-susceptible gene)를 탐색하는 연구가 매우 활발히 진행되고 있으며 더불어 효율적인 연구를 위한 연구설계의 중요성도 점차 증가하고 있다 (Terwilliger과 Ott, 1994). 질병 관련 유전자와 질병과의 연관성 분석(association analysis)을 위한 방법으로 환자-대조군 연구설계(case-control design)가 많이 사용되고 있다. 그 이유는 인구집단의 오랜 유전적 변이에 대한 정보가 대립유전자(allele)의 불평형(disequilibrium)으로 나타나는데 이를 비교함으로서 좁은 영역 내에서 질병 관련 유전자의 위치를 탐색할 수 있기 때문이다.

단일염기다형성(single nucleotide polymorphism; SNP)에 기초한 환자-대조군 연관성 분석에서는 환자군과 대조군의 대립유전자나 유전자형(genotype)의 빈도를 비교하는 방법을 사용한다 (Sasieni, 1997; Nielsen과 Weir, 1999). 그러나 인구집단이 층화(population stratification)되어 있으면 후보유

이 논문는 2008년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2008-313-C00150).
[4]교신저자: (120-752) 서울시 서대문구 신촌동 134번지 연세대학교 의과대학 예방의학교실, 교수.
E-mail: cmnam@yuhs.ac

전자와 질병 간에 아무런 관계가 없음에도 불구하고 연관되어 있다고 추론할 위양성(false positive)이 커지게 된다 (Setakis 등, 2006). 이와 같은 인구집단 층화를 해결하기 위해 대표적으로 Genomic control 방법 (Devlin과 Roeder, 1999)과 Structured association 방법 (Pritchard 등, 2000a, 2000b; Satten 등, 2001; Zhu 등, 2002; Hoggart 등, 2003)이 사용되고 있다.

SNP는 유전체(genome) 상에서 랜덤하게 분포하는 것이 아니라 블록(block)의 형태로 존재하기 때문에 단일 SNP 보다는 동일 유전자 상에서 서로 밀접하게 연결되어 있는 SNP들로 일배체형(haplotype)을 구성하고 연관불평형(linkage disequilibrium; LD)을 평가하는 것이 관련 유전자를 탐색하는데 유용한 것으로 평가되고 있다 (Jorde, 1995; Keavney, 2002).

그러나 SNP와 달리 일배체형에 기초한 연관성 분석에서는 유전자형이 unphased 상태로 주어지기 때문에 모든 SNP들이 동형(homozygous)이거나 임의의 한 SNP만 이형(heterozygous)인 경우가 아니면 일배체형 쌍이 유일하게 결정되지 않는다. 이와 같은 일배체형 쌍의 모호성(ambiguity)을 해결하기 위해 molecular haplotyping 기술을 적용할 수도 있지만 이 방법은 많은 비용과 시간이 필요하기 때문에 흔히 Hardy-Weinberg 평형(Hardy-Weinberg equilibrium; HWE) 가정 아래 EM(expectation-maximization) 알고리즘으로부터 일배체형 빈도를 추정하는 방법이 사용되고 있다 (Excoffier과 Slatkin, 1995; Long 등, 1995).

환자-대조군 연구에서 일배체형에 기초한 연관성 분석을 위해 Zhao 등 (2000)과 Fallin 등 (2001)은 EM 방법을 써서 각 군별로 일배체형의 빈도를 추정한 후 두 군간 일배체형 빈도를 비교하는 검정법을 제안하였다. 그러나 이 검정법은 일배체형과 질병의 연관성을 통합적으로 검정하기 때문에 특정 일배체형이 질병에 미치는 영향은 추론할 수 없었다. 그래서 Schaid 등 (2002)과 Zaykin 등 (2002)은 각 개체의 유전자형에 대응하는 일배체형이 주어졌을 때 질병 유무에 대한 전향적 우도(prospective likelihood)를 정의하여 이 문제를 해결하였다. 특히 Zaykin 등 (2002)은 각 개체의 모든 가능한 일배체형 쌍에 대한 사후확률(posterior probability)을 계산하여 일배체형들의 빈도를 추정한 후, 그 값을 독립변수로 하는 일배체형 경향성 회귀분석(haplotype trend regression; HTR)을 통해 특정 일배체형이 질병에 미치는 효과를 추정하였다. Epstein과 Satten (2003)은 질병이 주어졌을 때 일배체형에 대한 후향적 우도(retrospective likelihood)를 정의하고 여러 유전모형에 적용할 수 있는 방법을 제안하였다. Tanck 등 (2003)은 가중값으로 사후확률을 갖는 우도에, 빈도가 낮은 일배체형의 효과에 대한 추정이 안정되도록 하기 위해 두 일배체형에서 공유하는 대립유전자의 개수에 의존하는 벌점(penalty)을 추가하여 벌점 우도(penalized likelihood)를 정의하고, EM 방법을 써서 모수를 추정하였다.

일배체형에 기초한 여러 환자-대조군 연관성 분석 방법 중에서 Zaykin 등 (2002)의 HTR 방법은 SAS/Genetics (SAS Institute, 2002), R (http://www.R-project.org) 등과 같은 상용 소프트웨어에서 쉽게 이용할 수 있다. 본 연구에서는 모의실험 연구를 통해 인구집단 층화 및 인구집단에 대한 관측 오차(measurement error)가 HTR에 기초한 연관성 연구에 미치는 영향을 알아보고자 한다. 2절에서는 Zaykin 등 (2002)의 HTR 모형을 정리했으며, 인구집단 층화를 포함하는 확장된 HTR 모형을 제안하였다. 3절에서는 모의실험 연구를 통해 인구집단 층화가 단일 SNP 및 일배체형에 기초한 환자-대조군 연관성 검정의 제1종 오류율에 어떻게 영향을 주는 지를 살펴보았다. 4절에서는 환자-대조군 연관성 연구에 대해 고찰 하였으며 본 연구의 한계점 및 향후 과제를 제시하였다.

2. HTR 모형

2.1. Zaykin 등 (2002)의 모형

서로 독립인 n명의 개체가 있고, 각 개체는 연속된 $L(>0)$개의 SNP에 대한 유전자형을 갖는다고 가

정하자. 여기서 대조군과 환자군의 개체 수는 각각 c, d명이며, 따라서 $n = c + d$이다. 만일 모든 좌위(locus)에서 유전자형이 관측되면, 모든 가능한 유전자형과 일배체형은 각각 총 3^L, $2^L(\equiv m)$개이다. G는 어떤 개체의 유전자형을 나타내고, $H = (h, h')$는 그 개체에 대응하는 일배체형 쌍을 나타낸다고 하자. 그런데 어떤 개체가 이형인 좌위를 2개 이상 가지면 그 개체의 일배체형 쌍은 유일하게 결정되지 않는다. 유전자형 G에 대응하는 모든 가능한 일배체형 쌍들의 집합을 $\mathcal{S}(G)$라고 하자. 가령 $L = 3$이고 좌위 순서대로 SNP의 대립유전자가 $A|a$, $B|b$, $C|c$라고 하면 가능한 유전자형은 총 27개 이다. 그 중에서 유전자형 $G_1 = (AA, BB, CC)$, $G_2 = (AA, BB, Cc)$, $G_3 = (AA, Bb, Cc)$, $G_4 = (Aa, Bb, Cc)$ 만을 예로 들어 설명하면 각각 $\mathcal{S}(G_1) = \{H_1 = (ABC, ABC)\}$, $\mathcal{S}(G_2) = \{H_1 = (ABC, ABc)\}$, $\mathcal{S}(G_3) = \{H_1 = (ABC, Abc), H_2 = (ABc, AbC)\}$, $\mathcal{S}(G_4) = \{H_1 = (ABC, abc), H_2 = (ABc, abC), H_3 = (AbC, aBc), H_4 = (Abc, aBC)\}$와 같다. G_1, G_2에 대응하는 일배체형 쌍은 유일하지만, G_3, G_4에 대응하는 일배체형 쌍은 유일하지 않으며 조건부 확률의 크기에 따라 각 쌍이 확률적으로만 결정된다.

$Y_i (i = 1, \ldots, n)$는 i번째 개체의 질병 유무에 따라 각각 1 또는 0의 값을 갖는 이항반응변수라고 하자. $D_{ij} (i = 1, \ldots, n;\ j = 1, \ldots, m)$는 i번째 개체가 갖고 있는 j번째 일배체형의 빈도라고 하고, $\mathbf{D}_i' = (D_1, \ldots, D_{im}) (i = 1, \ldots, n)$를 i번째 개체 벡터라고 하자. 상술한 예에서는 $m = 8$이고, 일배체형들은 각각 $h_1 = ABC$, $h_2 = ABc$, $h_3 = AbC$, $h_4 = Abc$, $h_5 = aBC$, $h_6 = aBc$, $h_7 = abC$, $h_8 = abc$이다. $\mathcal{S}(G_3)$에 포함된 일배체형 쌍들의 조건부 확률을 $p_{3j} = \Pr(H_j|G_3) (j = 1, 2)$라고 하고, $\mathcal{S}(G_4)$에 포함된 일배체형 쌍들의 조건부 확률을 $p_{4j} = \Pr(H_j|G_4) (j = 1, \ldots, 4)$라고 하자. 여기서 $\sum_{j=1}^{2} p_{3j} = 1$, $\sum_{j=1}^{4} p_{4j} = 1$이다. 따라서 $\mathbf{D}_1'$ =(1, 0, 0, 0, 0, 0, 0, 0), $\mathbf{D}_2'$ =0.5(1, 1, 0, 0, 0, 0, 0, 0), $\mathbf{D}_3' = 0.5(p_{31}, p_{31}, p_{32}, p_{32}, 0, 0, 0, 0)$, $\mathbf{D}_4' = 0.5(p_{41}, p_{41}, p_{42}, p_{42}, p_{43}, p_{43}, p_{44}, p_{44})$이다. 일반적으로 유전자형이 $G_i (i = 1, \ldots, n)$인 개체가 일배체형 쌍 $H_j = (h_j, h_j') (j = 1, \ldots, k_i)$을 가질 조건부 확률(혹은 사후확률)은 HWE 가정 아래 다음과 같이 주어진다. 여기서 k_i는 집합 $\mathcal{S}(G_i)$의 크기를 나타낸다.

$$p_{j|i} = \Pr(H_j = (h_j, h_j')|G_i) = \frac{p_{h_j} p_{h_j'}}{\sum_{u=1}^{m} \sum_{v=1}^{m} \Pr(G_i|(h_u, h_v)) p_{h_u} p_{h_v}}, \quad i = 1, \ldots, n;\ j = 1, \ldots, k_i,$$

여기서 $p_h (h = h_1, \ldots, h_m)$는 일배체형 h의 빈도이며, $\Pr(G_i|(h_u, h_v))$은 일배체형 쌍 (h_u, h_v)가 유전자형 G_i에 대응하면 1이고, 그렇지 않으면 0이다. Zaykin 등 (2002)은 HTR 모형을 아래와 같이 정의하였다.

$$\text{Logit}\{\Pr(Y_i = 1)\} = \log\left\{\frac{\Pr(Y_i = 1)}{1 - \Pr(Y_i = 1)}\right\} = \beta_0 + \mathbf{D}_i'\beta, \quad i = 1, \ldots, n, \tag{2.1}$$

여기서 β_0는 절편항이고, $\beta' = (\beta_1, \ldots, \beta_m)$는 일배체형 효과에 대한 회귀계수 벡터이다. Zaykin 등 (2002)은 일배체형의 유의성, 즉 가설

$$H_0 : \beta_1 = \cdots = \beta_m = 0$$

을 우도비 검정(likelihood ratio test; LRT)을 통해 검정하였다.

2.2. 인구집단 층화를 포함하는 확장된 HTR 모형

본 연구에서는 인구집단 층화가 존재할 때 이를 해결하기 위한 방법으로 확장된 HTR(extended HTR; eHTR) 모형을 제안하고자 한다. 인구집단이 $s(> 1)$개의 층으로 층화되어 있다고 할 때, 가변수 I_{ik}를

다음과 같이 정의하자.

$$I_{ik} = \begin{cases} 1, & i \in P_k, \\ 0, & i \notin P_k, \end{cases} \quad i = 1, \ldots, n;\ k = 1, \ldots, s-1,$$

여기서 P_k은 인구집단 k에 속하는 개체들로 이루어진 집합을 나타낸다. $\mathbf{I}_i' = (I_{i1}, \ldots, I_{i,s-1})$는 i번째 개체의 인구집단 층화 벡터라고 하자. eHTR 모형을 아래와 같이 정의하자.

$$\text{Logit}\{\Pr(Y_i = 1)\} = \beta_0 + \mathbf{D}_i'\beta + \mathbf{I}_i'\gamma, \quad i = 1, \ldots, n, \tag{2.2}$$

여기서 $\gamma' = (\gamma_1, \ldots, \gamma_{s-1})$는 인구집단 층화 효과에 대한 회귀계수 벡터이다. 모형 (2.2)는 각 개체가 속하는 인구집단을 안다는 가정 아래 인구집단의 층화에 따른 교락효과(confounding effect)를 제어하기 위해 인구집단 층화에 대한 정보를 모형 (2.1)에 포함시킨 것이다. 인구집단 층화에 따른 교락효과를 제어한 후 Zaykin 등 (2002)처럼 일배체체형이 질병에 미치는 효과에 대한 추정뿐만 아니라 LRT을 통해 일배체형의 유의성에 대한 가설 H_0를 검정할 수 있다.

3. 모의실험 연구

3.1. 분류 오류와 자료 생성

인구집단 층화가 가설 $H_0 : \beta_1 = \cdots = \beta_m = 0$의 검정에 미치는 영향을 알아보기 위해 모의실험 연구를 수행하였다. 이를 위해 두 모형 (2.1)과 (2.2)에서 제1종 오류율이 얼마나 잘 조절되는지를 살펴보았다. 한편 각 개체가 속하는 인구집단을 알지 못할 때는 그 개체가 속할 인구집단을 추정해야 하는데 이때 추정에 따른 분류 오류(misclassification error)가 발생할 수 있다. 이와 같은 분류 오류를 모의실험에 반영하기 위해 민감도(sensitivity; Se)와 특이도(specificity; Sp)를 고려하였다. 특히 인구집단 층이 2개일 때, 즉 $s = 2$일 때, 민감도는 참(true) 인구집단 1에 속한 개체를 인구집단 1로 분류하는 비율을 의미하며, 특이도는 참 인구집단 2에 속한 개체를 인구집단 2로 분류하는 비율을 의미한다.

모의실험을 위한 자료는 Kim 등 (2004)과 유사하게 아래와 같이 생성하였다.

- 일배체형 블록에 3개의 표지유전자 좌위가 있고, 각각의 표지유전자 좌위에는 2개의 대립유전자 1과 2를 갖는다고 가정하였다($L = 3, m = 8$). 따라서 8개 일배체형은 각각 $h_1 = 111$, $h_2 = 112$, $h_3 = 121$, $h_4 = 122$, $h_5 = 211$, $h_6 = 212$, $h_7 = 221$, $h_8 = 222$이다.
- 인구집단 층이 2개라고 가정하였다($s = 2$).
- 각 층의 일배체형 빈도는 표 3.1과 같이 가정하였다. 두 인구집단의 일배체형 분포의 차이는 q의 크기에 따라 결정되는데, $q = 0.1$이면 두 분포는 동일하고, q의 값이 커지면 두 분포의 차이도 커진다. $q = 0.1, 0.2, 0.3, 0.4$를 각각 가정하였다.
- 각 개체의 유전자형은 해당 개체가 속한 인구집단으로부터 생성한 2개의 일배체형을 묶어 정의하였다. 따라서 가능한 유전자형은 총 27개이며, 각각 $G_1 = (11, 11, 11)$, $G_2 = (11, 11, 12)$, $G_3 = (11, 11, 22)$, $G_4 = (11, 12, 11)$, $G_5 = (11, 12, 12)$, $G_6 = (11, 12, 22)$, $G_7 = (11, 22, 11)$, $G_8 = (11, 12, 12)$, $G_9 = (11, 12, 22)$, ...와 같다.
- 모의실험에서는 제1종 오류율만을 검토하기 때문에 모든 일배체형의 질병 발생 위험이 동일하다고 가정하였다.
- 인구집단 2의 질병 발생 위험이 인구집단 1보다 RR $= 1.0, 2.0, 3.0$배 높다고 각각 가정하였다.

표 **3.1.** 인구집단별 일배체형의 분포, $q = 0.1, 0.2, 0.3, 0.4$

인구집단	일배체형							
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
1	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{10}$	$\frac{1}{10}$
2	$\frac{1-2q}{6}$	$\frac{1-2q}{6}$	$\frac{1-2q}{6}$	$\frac{1-2q}{6}$	$\frac{1-2q}{6}$	$\frac{1-2q}{6}$	q	q

- 민감도(Se)와 특이도(Sp)는 동일하고, Se(Sp) = 1.0(분류 오류가 없는 경우); 0.95, 0.9, 0.8(분류 오류가 있는 경우)를 각각 가정하였다.
- 환자군과 대조군은 각각 500명을 대상으로 하였다($c = d = 500$, $n = 1{,}000$).
- 총 1,000개의 자료집합을 생성하였다.

3.2. 경험적 유의확률의 추정

3.1절에서 생성한 각각의 자료집합에 대해 아래와 같은 과정을 1,000번 반복하여 유의수준 5%에서 경험적 유의확률을 추정하였다.

단계 1: 환자군과 대조군을 합친 유전자형 자료로부터 EM 알고리즘을 이용하여 일배체형 빈도를 추정한다($\hat{p}_h$, $h = h_1, \ldots, h_8$).

단계 2: 각 개체의 유전자형 $G_i (i = 1, \ldots, n)$으로부터 모든 가능한 일배체형 쌍 H_j의 사후확률을 추정한다($\hat{p}_{j|i}, i = 1, \ldots, n;\ j = 1, \ldots, k_i$). 일배체형 쌍 H_j에 대응하는 일배체형 h_j와 h'_j의 빈도를 $0.5 \times \hat{p}_{j|i}$으로 추정하여 벡터 $\mathbf{D}'_i$를 추정한다($\hat{\mathbf{D}}'_i$, $i = 1, \ldots, n$).

단계 3: 민감도(혹은 특이도)에 따라 각 개체가 속할 인구집단으로 분류한다. 분류된 인구집단으로 벡터 $\mathbf{I}_i$를 추정한다($\hat{\mathbf{I}}_i, i = 1, \ldots, n$). 따라서 Se(혹은 Sp) = 1이면, $\hat{\mathbf{I}}_i = \mathbf{I}_i$을 항상 만족하지만, 그 외의 경우(즉, 0.95, 0.9, 0.8)에는 항상 만족되는 것은 아니다. 가령 Se(Sp) = $c \in (0, 1)$일 때, 균등분포 $U(0, 1)$에서 난수를 발생하여 그 값이 c보다 작으면 인구집단 1에 속한 개체는 1, 인구집단 2에 속한 개체는 2로 분류하고(즉, 분류 오류가 없음), 그렇지 않으면 인구집단 1에 속한 개체는 2, 인구집단 2에 속한 개체는 1로 분류한다(즉, 분류 오류가 있음).

단계 4: 모형 (2.1)과 (2.2)에서 가설 H_0를 검정하기 위해 각각 모형 (2.1)는 아래 모형 (3.1)와 비교하고, 모형 (2.2)는 아래 모형 (3.2)와 비교한다. 전자는 검정통계량 G_1^2(모형(2.1)|모형(3.1)) $= -2(ll(\text{모형}(3.1)) - ll(\text{모형}(2.1)))$을 이용하고, 후자는 검정통계량 G_2^2(모형(2.2)|모형(3.2)) $= -2(ll(\text{모형}(3.2)) - ll(\text{모형}(2.2)))$을 이용한다. 여기서 ll은 각 모형의 최대 로그우도 값을 나타낸다. G_1^2와 G_2^2는 각각 근사적으로 자유도 8인 카이제곱분포를 따르므로 이를 이용하여 가설 H_0를 검정한다.

$$\text{Logit}\{\Pr(Y_i = 1)\} = \beta_0, \quad i = 1, \ldots, n. \tag{3.1}$$

$$\text{Logit}\{\Pr(Y_i = 1)\} = \beta_0 + \mathbf{I}'_i\gamma, \quad i = 1, \ldots, n. \tag{3.2}$$

3.3. 모의실험 결과

그림 3.1은 두 인구집단에 따른 유전자 좌위별 대립유전자의 분포와 일배체형의 분포를 나타낸다. 특히 유전자 좌위별 대립유전자의 분포에서는 이대립유전자(biallele)를 고려하고 있기 때문에 'SNP = 1'의

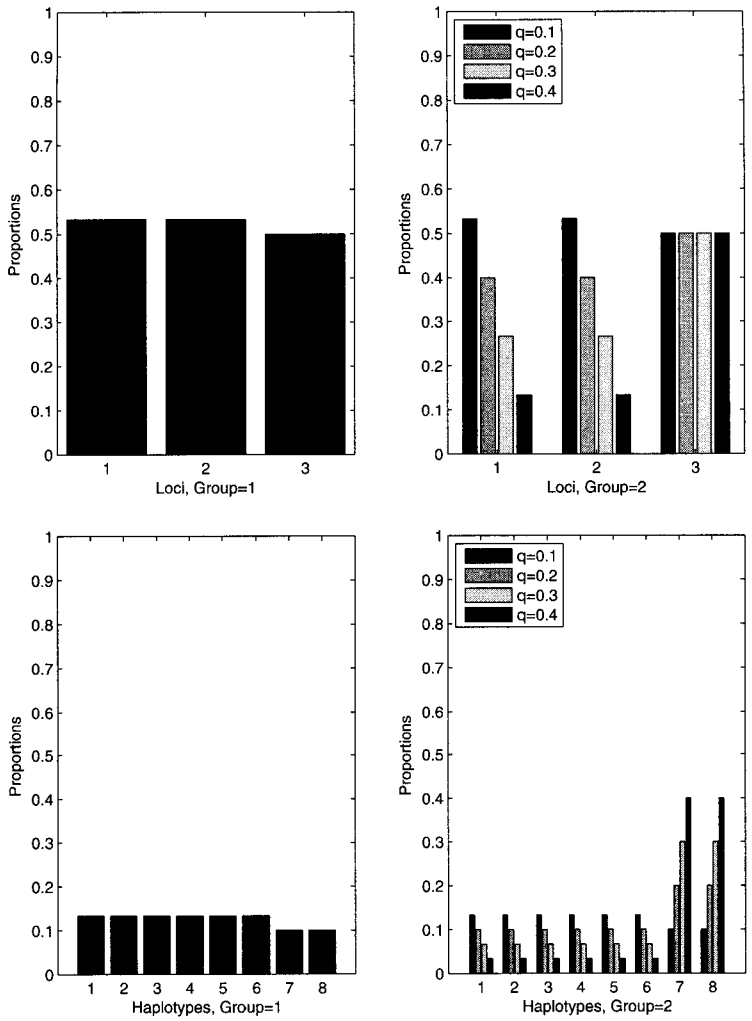

그림 **3.1.** 두 인구집단에 따른 유전자 좌위별 대립유전자 'SNP = 1'의 분포와 일배체형의 분포

분포만을 나타냈다. 유전자 좌위를 l로 표시하자. $l = 1$과 $l = 2$의 SNP들은 분포가 서로 같고, 두 인구집단 간 대립유전자 1(또는 2)의 비율을 비교해 보면 q의 값이 커질수록 인구집단 2의 비율이 인구집단 1(baseline)의 비율보다 점차 작아진다(또는 커진다). $l = 3$의 SNP는 q의 값에 관계 없이 두 인구 집단 간 대립유전자 1(또는 2)의 비율이 같다. 한편 일배체형 $h_1, h_2, h_3, h_4, h_5, h_6$의 분포는 q의 값이 커질수록 동일하게 인구집단 2의 비율이 인구집단 1의 비율보다 점차 작아지고, 일배체형 h_7과 h_8의 분포는 그와 반대로 점차 커진다.

표 3.2은 각 유전자 좌위에서 대립유전자의 유의성 검정을 위한 모형 (2.1)과 모형 (2.2)의 제1종 오류율을 1,000번 반복하여 얻은 결과이다. 그림 3.1에서 예상했던 것처럼 $l = 1$과 $l = 2$에 해당하는 결과들은 서로 유사한 경향을 띄고 있다. RR = 1일 때는 두 인구집단의 위험률이 같기 때문에 제1종 오류율이 비교적 잘 조절된다고 할 수 있다. 그러나 q의 값이 증가할수록 두 인구집단의 분포가 많이 상이해지기

표 3.2. 각 유전자 좌위에서 대립유전자의 유의성 검정을 위한 모형 (2.1) 'HTR'과 모형 (2.2) 'eHTR'의 제1종 오류율, 반복횟수 = 1,000

			q							
			0.1		0.2		0.3		0.4	
l	RR	Se(Sp)	HTR	eHTR	HTR	eHTR	HTR	eHTR	HTR	eHTR
1	1	1.00	0.051	0.051	0.038	0.043	0.040	0.047	0.017	0.048
		0.95	0.043	0.043	0.041	0.044	0.029	0.043	0.023	0.043
		0.90	0.037	0.040	0.048	0.051	0.034	0.048	0.017	0.046
		0.80	0.055	0.054	0.050	0.051	0.033	0.033	0.021	0.038
	2	1.00	0.043	0.041	0.159	0.040	0.491	0.053	0.887	0.046
		0.95	0.039	0.039	0.167	0.057	0.526	0.075	0.890	0.100
		0.90	0.044	0.044	0.146	0.067	0.514	0.126	0.890	0.232
		0.80	0.059	0.056	0.177	0.108	0.513	0.267	0.901	0.523
	3	1.00	0.049	0.052	0.337	0.046	0.888	0.053	0.999	0.046
		0.95	0.058	0.054	0.306	0.054	0.866	0.122	0.999	0.184
		0.90	0.063	0.062	0.300	0.089	0.861	0.231	0.998	0.497
		0.80	0.058	0.056	0.301	0.177	0.863	0.551	0.999	0.896
2	1	1.00	0.052	0.052	0.047	0.055	0.031	0.042	0.013	0.035
		0.95	0.044	0.045	0.046	0.054	0.032	0.050	0.015	0.039
		0.90	0.050	0.050	0.049	0.053	0.032	0.048	0.016	0.044
		0.80	0.055	0.056	0.042	0.042	0.041	0.049	0.017	0.033
	2	1.00	0.053	0.053	0.168	0.044	0.532	0.056	0.904	0.050
		0.95	0.046	0.042	0.162	0.066	0.504	0.057	0.911	0.099
		0.90	0.060	0.064	0.160	0.069	0.568	0.127	0.890	0.240
		0.80	0.043	0.046	0.170	0.096	0.500	0.263	0.913	0.545
	3	1.00	0.049	0.046	0.302	0.058	0.865	0.042	0.999	0.060
		0.95	0.045	0.047	0.332	0.069	0.866	0.111	0.998	0.192
		0.90	0.066	0.059	0.324	0.082	0.873	0.212	0.999	0.491
		0.80	0.049	0.055	0.309	0.179	0.842	0.487	1.000	0.901
3	1	1.00	0.052	0.052	0.052	0.052	0.049	0.049	0.059	0.059
		0.95	0.047	0.047	0.058	0.057	0.059	0.058	0.037	0.038
		0.90	0.042	0.042	0.052	0.052	0.042	0.043	0.057	0.059
		0.80	0.039	0.040	0.044	0.043	0.050	0.050	0.050	0.049
	2	1.00	0.048	0.056	0.041	0.042	0.053	0.053	0.042	0.043
		0.95	0.057	0.055	0.048	0.047	0.045	0.044	0.039	0.039
		0.90	0.042	0.044	0.056	0.051	0.048	0.044	0.049	0.045
		0.80	0.050	0.047	0.049	0.047	0.048	0.048	0.059	0.059
	3	1.00	0.042	0.048	0.058	0.062	0.047	0.046	0.041	0.047
		0.95	0.051	0.051	0.048	0.049	0.049	0.041	0.048	0.043
		0.90	0.054	0.057	0.047	0.052	0.051	0.051	0.048	0.057
		0.80	0.051	0.050	0.041	0.040	0.055	0.057	0.060	0.062

때문에 이에 대해 보정을 하지 않은 모형 (2.1)은 점차 보수적인 검정이 된다. 이에 비해 모형 (2.2)는 비교적 덜 보수적이지만 Se(Sp)의 값이 작아지면 분류 오류가 점차 늘어나기 때문에 이로 인해 약간 보수적인 검정이 된다. RR = 2, 3일 때는 RR = 1일 때와 매우 다른 특징을 찾아볼 수 있다. RR이 커짐에 따라 두 집단의 질병 발생 비율이 매우 달라지기 때문에 대립유전자의 유의성 검정에서 인구집단의 층화에 따른 위양성이 나타날 수 있다. 실제로 모형 (2.1)은 $q = 0.2$일 때 제1종 오류율이 15–18%(RR

표 **3.3.** 일배체형의 유의성 검정을 위한 모형 (2.1) 'HTR'과 모형 (2.2) 'eHTR'의 제1종 오류율, 반복횟수 = 1,000

		q							
		0.1		0.2		0.3		0.4	
RR	Se(Sp)	HTR	eHTR	HTR	eHTR	HTR	eHTR	HTR	eHTR
1	1.00	0.056	0.056	0.041	0.044	0.043	0.054	0.036	0.054
	0.95	0.057	0.060	0.048	0.048	0.062	0.066	0.040	0.052
	0.90	0.046	0.046	0.050	0.052	0.042	0.047	0.045	0.054
	0.80	0.052	0.054	0.041	0.044	0.032	0.038	0.044	0.047
2	1.00	0.048	0.048	0.125	0.037	0.427	0.048	0.859	0.064
	0.95	0.053	0.050	0.109	0.051	0.418	0.057	0.848	0.095
	0.90	0.056	0.057	0.154	0.068	0.424	0.106	0.833	0.163
	0.80	0.045	0.042	0.140	0.080	0.429	0.196	0.841	0.429
3	1.00	0.044	0.043	0.281	0.049	0.859	0.056	0.999	0.062
	0.95	0.055	0.054	0.296	0.076	0.880	0.097	0.999	0.156
	0.90	0.054	0.061	0.276	0.106	0.862	0.171	1.000	0.390
	0.80	0.057	0.058	0.297	0.153	0.864	0.449	0.999	0.880

= 2), 30–34%(RR = 3) 정도이다. 그런데 $q = 0.3, 0.4$일 때는 제1종 오류율이 이보다 훨씬 커졌다. 모형 (2.2)도 모형 (2.1)과 마찬가지로 제1종 오류율이 조절되지 못했지만 모형 (2.1)의 제1종 오류율보다는 훨씬 작았다. 인구집단 층화를 제어 했음에도 불구하고 모형 (2.2)에서 제1종 오류율이 조절되지 못하는 것은 Se(Sp)의 값이 작아짐에 따라 분류 오류가 점차 늘어나기 때문이다. 이 결과로부터 알 수 있는 것은 SNP에 기초한 환자-대조군 연관성 연구에서 인구집단 층화에 대한 정보를 갖고 있다할지라도 그 속에 관측 오차가 존재하면, 즉 민감도(특이도)가 1이 아니면, 제1종 오류율이 조절되지 않는다는 것이다. 실제로는 가능하지 않겠지만 인구집단 층화에 대한 관측 오차가 전혀 없으면, 즉 민감도(특이도)가 1이면, q의 값에 관계 없이 모형 (2.2)는 제1종 오류율이 잘 조절되는 것으로 나타났다. $l = 3$에 있는 대립유전자는 q의 값에 관계 없이 두 인구집단의 분포가 정확하게 같기 때문에 Sp(Se)의 값에 제1종 오류율이 영향을 받지 않으며, 또한 RR이 1보다 큰 값을 갖는다할지라도 H_0 하에서는 대립유전자와 질병이 연관되어 있지 않기 때문에 두 모형 모두 제1종 오류율이 잘 조절되었다.

표 3.3은 일배체형의 유의성 검정을 위한 모형 (2.1)과 (2.2)의 제1종 오류율을 1,000번 반복하여 얻은 결과이다. 모형 (2.1)은 인구집단 층화가 없거나($q = 0.1$) 또는 인구집단 층화가 존재해도 두 인구집단의 질병 발생 비율이 같으면(RR = 1) 제1종 오류율을 잘 조절하는 것으로 나타났다. 그러나 두 인구집단의 질병 발생 비율이 서로 다를 때는(RR = 2 또는 3) 일배체형 분포의 차이가 커질수록(즉, q의 값이 커질수록) 제1종 오류율이 점차 조절되지 않는 것으로 나타났다. 실제로 $q = 0.2$일 때는 제1종 오류율이 11–15%(RR = 2), 28–30%(RR = 3), $q = 0.3$일 때는 42–43%(RR = 2), 86–88%(RR = 3), $q = 0.4$일 때는 83–86%(RR = 2), 100%(RR = 3)로 각각 나타났다. 따라서 모형 (2.1)에 기초한 일배체형의 유의성 검정은 일배체형이 질병과 연관성이 없음에도 불구하고 인구집단 층화가 뚜렷하게 존재할 때 제1종 오류율이 명목 유의수준 보다 매우 커지는 문제를 제어할 수 없었다. 한편 모형 (2.2)도 모형 (2.1)과 마찬가지로 인구집단 층화가 없거나($q = 0.1$) 또는 인구집단 층화가 존재해도 두 인구집단의 질병 발생 비율이 같으면(RR = 1) 제1종 오류율을 잘 조절할 수 있었다. 또한 두 인구집단의 질병 발생 비율이 다르고(RR = 2 또는 3) 일배체형 분포가 다르다 할지라도($q = 0.2$, 0.3 또는 0.4) 민감도(특이도)가 1이면(즉, 분류 오류가 없으면) 모형 (2.1)과 달리 제1종 오류율이 잘 조절되었다. 그러나 두 인구집단의 질병 발생 비율이 다르고 일배체형 분포가 다를 때 분류 오류가 존재하면(Se(Sp) = 0.95, 0.9, 0.8) 제1종 오류율이 잘 조절되지는 않았지만 모형 (2.1)의 제1종 오류율 보다는 훨씬 작게 나타났

다. 표 3.2의 SNP에 기초한 환자-대조군 연관성 연구에서 살펴본 것처럼 일배체형에 기초한 환자-대조군 연관성 연구에서도 인구집단 층화에 대한 정보를 갖고 있다할지라도 그 속에 관측 오차가 존재하면 위양성을 피할 수는 없었다. 그러나 인구집단 층화에 대한 관측 오차가 전혀 없으면 인구집단 간 질병 발생 비율이나 일배체형의 분포가 서로 다르다 할지라도 이에 관계 없이 모형 (2.2)는 제1종 오류율을 잘 만족하였다.

4. 고찰

새로운 유전자의 위치를 탐색하는 방법으로서 환자-대조군을 이용한 연관성 분석은 자료를 수집하거나 통계적 분석방법이 용이하고, 또한 유전체 상에서 관련 유전자의 위치를 좁은 영역에서 탐색하는 것이 가능하기 때문에 많이 사용하고 있다. 그러나 환자-대조군 연관성 연구는 인구집단 층화로 야기되는 위양성과 같은 문제점을 갖고 있다. 표 3.2에서 살펴본 것처럼 후보 유전자가 질병과 아무런 관계가 없다 할지라도 혼합된 인구집단의 질병 발생 비율이나 대립유전자의 분포가 상이할수록 제1종 오류율이 명목 유의수준 보다 커지는 경향이 있었다.

SNP에 기초한 환자-대조군 연관성 연구에서 인구집단 층화로 인한 위양성을 해결하기 위해 Devlin과 Roeder (1999)가 제안한 Genomic control(GC) 방법과 Pritchard 등 (2000a, 2000b), Satten 등 (2001), Zhu 등 (2002), Hoggart 등 (2003)이 제안한 Structured association(SA) 방법이 널리 사용되고 있다. 질병과 연관성이 없는 표지자에 대한 Armitage의 경향성 검정 (Armitage, 1955)은 근사적으로 자유도가 1인 카이제곱 분포를 따라야 하는데, 인구집단 층화가 있으면 검정통계량의 값이 명목값 보다 커진다. Devlin과 Roeder (1999)는 이 분산팽창계수(variance inflation factor)을 추정하여 Armitage의 경향성 검정통계량을 수정하는 방법을 제안하였다. SA는 GC와 마찬가지로 관련 질환과 연관성이 없는 많은 표지자들를 이용하여 연구 대상자들의 인구집단을 추정하고, 각 인구집단 내에서 연관성 분석을 시행하거나 또는 그 결과들을 통계적으로 병합하는 방법이다 (Pritchard 등, 2000a, 2000b). Satten 등 (2001)은 사회과학 분야에서 유사한 반응 행태를 보이는 개체들을 그룹화 하기 위해 널리 쓰이고 있는 Latent-class 방법을 제안하였다. 실제로는 몇 개의 인구집단으로 층화되어 있는지를 모르기 때문에 Zhu 등 (2002)은 그 개수를 추정하는 방법을 제안하였다. Hoggart 등 (2003)은 기존의 방법과 베이지안 방법을 결합하는 방법을 제안하였다.

한편 일배체형을 이용한 환자-대조군 연관성 분석에서 가장 널리 사용되는 방법은 SAS/Genetic (SAS Institute, 2002)와 R (http://www.R-project.org) 등과 같은 상용소프트웨어에서 쉽게 이용할 수 있는 Zaykin 등 (2002)이 제안한 HTR 방법이다. 표 3.3에서 살펴본 것처럼 HTR 모형은 인구집단 층화에 민감한 단점을 가지고 있다. 본 연구에서는 이를 해결하기 위한 한 방법으로 HTR 모형에 인구집단의 정보를 추가하여 eHTR 모형을 제안하였다. 비록 2절에서는 이 대립유전자를 중심으로 두 가지 검정방법을 소개했지만 다중 대립유전자(multiple allele)을 갖는 경우로도 확장할 수 있다. 가령 임의의 한 유전자 좌위에서 세 종류의 대립유전자 A, B, C를 갖는다고 하자. 이 때, 가능한 유전자형은 6가지이고(즉, AA, AB, AC, BB, BC, CC), 가능한 일배체형은 3가지이다(즉, A, B, C). 각 유전자형에 대응하는 일배체형 쌍은 유일하게 결정되며, 따라서 유전자형이 'AA'이면 모형 (2.1)과 모형 (2.2)에 포함된 개체벡터는 $\mathbf{D}' = (1,0,0)$으로 정의되고, 유전자형이 'AB', 'AC', 'BB', 'BC', 'CC'이면 각각 $\mathbf{D}' = (0.5,0.5,0)$, $\mathbf{D}' = (0.5,0,0.5)$, $\mathbf{D}' = (0,1,0)$, $\mathbf{D}' = (0,0.5,0.5)$, $\mathbf{D}' = (0,0,1)$으로 정의된다. 또한, 2개 이상의 서로 다른 유전자 좌위에 있는 다중 대립유전자를 이용한 일배체형에 대해서도 동일한 방법으로 확장이 가능하다. 모의실험 결과에 의하면 인구집단 층화가 있지만 참 인구집단을 알고 있다고 가정할 때 HTR 모형은 제1종 오류율을 잘 조절하지 못했지만 본 연구에서 제안한 eHTR 모형은 제1종 오류율을 잘 조절하는 것으로 나타났다. 그러나 민감도(특이도)가 작아짐에 따라(즉, 분류 오류

가 커짐에 따라) eHTR 모형도 제1종 오류율이 명목 유의수준 보다 커지는 경향이 있었다. 결국 인구집단 층화에 대한 정보를 갖고 있다할지라도 그 속에 관측 오차가 존재하면 위양성을 피하기 어렵다는 것을 알 수 있었다. 일반적으로 민감도와 특이도는 교역하는 성질이 있기 때문에 민감도를 특이도 보다 크게 하거나 특이도를 민감도 보다 크게 할 수 있지만 이에 따른 여러 조합이 가능하여 본 연구에서는 두 값이 모두 같은 경우만을 고려하였다. 한편, 본 연구에서는 인구집단 층화가 일배체형의 유의성 검정에 미치는 영향을 밝히는 데 목적을 두고 있지만 제1종 오류율이 조절되는 제한된 조건하에서 두 모형의 검정력을 비교하였다. 이를 위해 3.1에서 소개한 일배체형 중에서 h_7과 h_8은 고위험(high risk) 일배체형으로 분류하고, 그 외 일배체형은 저위험(low risk) 일배체형으로 분류하였다. 대조군의 일배체형 분포는 표 3.1에 있는 인구집단 1의 분포와 같이 정의하였고, 환자군의 일배체형의 분포는 오즈비(odds ratio; OR)를 고려하여 결정하였다. 모의실험에 포함된 OR은 1, 1.2, 1.65이며, 'OR = 1'는 귀무가설에 해당되고, 그 외 경우는 환자군에서 고위험 일배체형을 가질 오즈가 대조군 보다 각각 1.2배, 1.65배 높은 경우에 해당된다. 제1종 오류율이 조절되는 조건 하에서 모의실험을 하기 위해 인구집단 1과 2의 대조군과 환자군의 분포를 같게 하였고, 1,000번 모의실험을 수행 본 결과 예상했던 것처럼 두 모형의 검정력은 거의 같은 것으로 나타났다('OR = 1.2'이면, 검정력 = 31–32%, 'OR=1.65'이면 검정력 = 91–92%).

본 연구의 모의실험 결과를 통해 살펴보았듯이 환자-대조군 연관성 연구에서 인구집단 층화가 존재할 때 위양성이 없는 검정을 위해 필요한 것은 분류 오류를 줄여나가야 한다는 점이다. 향후 연구에서는 여러 유전모형에 대해 SA에서 제안된 방법들의 오류율을 비교하고자 하며 그 결과를 일배체형에 기초한 환자-대조군 연관성 연구에 활용하고자 한다.

참고문헌

Armitage, P. (1955). Tests for linear trends in proportions and frequencies, *Biometrics*, **11**, 375–386.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies, *Biometrics*, **55**, 997–1004.

Epstein, M. P. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data, *The American Journal of Human Genetics*, **73**, 1316–1329.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, *Molecular Biology and Evolution*, **12**, 921–927.

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D. and Schork, N. J. (2001). Genetic analysis of case/control data using estimated haplotype frequencies: Application to APOE locus variation and alzheimer's disease, *Genome Research*, **11**, 143–151.

Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G. and McKeigue, P. M. (2003). Control of confounding of genetic associations in stratified populations, *The American Journal of Human Genetics*, **72**, 1492–1504.

Jorde, L. B. (1995). Linkage disequilibrium as a gene-mapping tool, *The American Journal of Human Genetics*, **56**, 11–14.

Keavney, B. (2002). Genetic epidemiological studies of coronary heart disease, *International Journal of Epidemiology*, **31**, 730–736.

Kim, J., Kang, D. R., Lee, Y. K., Shin, S. M., Suh, I. and Nam, C. M. (2004). Statistical algorithm in genetic linkage based on haplotypes, *Journal of Preventive Medicine and Public Health*, **37**, 366–372.

Long, J. C., Williams, R. C. and Urbanek, M. (1995). An E-M algorithm and testing strategy for multiple-locus haplotypes, *The American Journal of Human Genetics*, **56**, 799–810.

Nielsen, D. M. and Weir, B. S. (1999). A classical setting for associations between markers and loci affecting quantitative traits, *Genetical Research*, **74**, 271–277.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000a). Inference of population structure using multilocus genotype data, *Genetics*, **155**, 945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000b). Association mapping in structured populations, *The American Journal of Human Genetics*, **67**, 170–181.

SAS Institute. (2002). *SAS/Genetics User's Guide*, SAS Institute, Cary.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size, *Biometrics*, **53**, 1253–1261.

Satten, G. A., Flanders, W. D. and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model, *The American Journal of Human Genetics*, **68**, 466–477.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous, *The American Journal of Human Genetics*, **70**, 425–434.

Setakis, E., Stirnadel, H. and Balding, D. J. (2006). Logistic regression protects against population structure in genetic association studies, *Genome Research*, **16**, 290–296.

Tanck, M. W. T., Klerkx, A. H. E. M., Jukema, J. W., De Knijff, P., Kastelein, J. J. P. and Zwinderman, A. H. (2003). Estimation of multilocus haplotype effects using weighted penalised log-likelihood: Analysis of five sequence variations at the cholesteryl ester transfer protein gene locus, *Annals of Human Genetics*, **67**, 175–184.

Terwilliger, J. and Ott, J. (1994). *Handbook of Human Genetic Linkage*, Johns Hopkins University Press, Baltimore.

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J. and Ehm, M. G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals, *Human Heredity*, **53**, 79–91.

Zhao, J. H., Curtis, D. and Sham, P. C. (2000). Model-free analysis and permutation tests for allelic associations, *Human Heredity*, **50**, 133–139.

Zhu, X., Zhang, S., Zhao, H. and Cooper, R. S. (2002). Association mapping, using a mixture model for complex traits, *Genetic Epidemiology*, **23**, 181–196.

Study on Effects of Population Stratification on Haplotype Trend Test in Case-Control Studies

Jinheum Kim[1] · Dae Ryong Kang[2] · Hyunsun Lim[3] · Chung Mo Nam[4]

[1]Department of Applied Statistics, University of Suwon
[2]Clinical Trials Center, Severance Hospital, Yonsei University
[3]Department of Biostatistics, Yonsei University College of Medicine
[4]Department of Preventive Medicine, Yonsei University College of Medicine

(Received August 2009; accepted September 2009)

Abstract

Population stratification can cause spurious associations between genetic markers and disease locus. In order to handle this population stratification in haplotype-based case-control association studies, we added population indicators as covariates to the haplotype trend regression model proposed by Zaykin *et al.* (2002). We investigated through simulations how both population stratification and measurement error in the estimation of true population of each individual affect type I error probabilities of the association tests based on both Zaykin *et al.*'s (2002) model and the proposed model. Based on those results, in the situation that there exists population stratification but there is no error in population classification of each individual, our proposed model does satisfy a type I error probability whereas Zaykin *et al.*'s (2002) model does not. However, as the measurement error increases, a type I error probability of our model correspondingly becomes larger than a nominal significance level. It implies that as long as uncertainty in the estimation of true population of each individual still remains, it is nearly impossible to avoid false positive in case-control association studies based on haplotypes.

Keywords: Population stratification, spurious association, false positive, haplotype trend test, measurement error.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD)(KRF-2008-313-C00150).

[4]Corresponding author: Associate Professor, Department of Preventive Medicine, Yonsei University College of Medicine, Seoul 120-749, Korea. E-mail: cmnam@yuhs.ac