응용통계연구 제 18권 2호, 2005년, pp. 297-310

일배체형에 기초한 고혈압과 ACE 유전자의 연관성 분석*

김진흠[1] 남정모 [2] 강대룡 [3] 서일 [4]

요 약

본 연구에서는 강화연구(서일, 2004)를 통해 수집된 277명의 환자-대조군 자료와 40개의 가계 자료를 이용하여 ACE(angiotensin-I converting enzyme) 유전자 내에 있는 4개의 단일염기다형성(single nucleotide polymorphism; SNP)으로 이루어진 일배체형 (haplotype)과 고혈압의 관련성을 알아보고자 한다. 이를 위해 일배체형에 기초한 환자-대조군의 우도비 검정과 가계 자료의 TDT(transmission/disequilibrium test) 연구를 수행하고자 한다. 또한 이 일배체형을 동정(identification)할 수 있는 tag-SNPs에 기초하여 동일한 연구를 하고자 하며, Seltman 등(2003)이 제안한 분기도(cladogram) 분석 방법을 써서 일배체형의 진화 과정에서 가깝게 위치하고 질병 발생 위험이 비슷한 클레이드(clade)를 찾아내고 이 클레이드와 고혈압의 연관성을 살펴보고자 한다.

주요용어: ACE 유전자; TDT; 가계 자료; 일배체형; 환자-대조군

1. 서론

최근 염기서열 분석 기술의 눈부신 성장과 국제간 SNP 협력 연구 및 HapMap 프로젝트 연구 결과 (Gabriel 등, 2002; The International HapMap Consortium, 2003)는 질병 관련 유전자의 위치를 탐색하는 방법에 큰 변화를 가져왔다. 지금 까지는 주로 유전자의 위치를 탐색하는데 하나의 표지 유전자(genetic markers) 또는 SNP를 이용하는 방법에 의존해왔지만, 앞으로는 유전체 상에 가깝게 위치한 SNP들의 묶음인 일배체형을 이용하는 방법에 크게 의존할 것으로 기대된다. 이러한 연구 방법의 변화는 외적인 성장과 더불어 단일 유전자 좌위(locus) 정보를 이용한 연쇄 분석(linkage analysis) 또는 연관성 분석(association analysis)보다는 일배체형의 정보를 이용한 분석이 통계적 검정력을 높일 수 있기 때문이다(Jorde, 1995; Keavney, 2002).

일배체형에 기초한 연쇄 분석과 연관성 분석을 위해 먼저 각 개체의 일배체형 정보를 알아야 한다. 그런데 일반적으로 우리에게 주어지는 자료는 일배체형 대신에 유전자

* 본 연구는 보건복지부 보건의료기술진흥사업의 지원(03-PJ1-PG3-21000-0015, 남정모)에 의하여 연구되었음.

1) (445-743) 경기도 화성시 봉담읍 와우리 산 2-2호, 수원대학교 자연과학대학 통계정보학과, 부교수
E-mail: jinhkim@suwon.ac.kr
2) (120-749) 서울시 서대문구 신촌동 134번지 연세대학교 의과대학 예방의학교실, 부교수
E-mail: cmnam@yumc.yonsei.ac.kr
3) (120-749) 서울시 서대문구 신촌동 134번지 연세대학교 의과대학 예방의학교실, 연구강사
E-mail: cosakang@yumc.yonsei.ac.kr
4) (120-749) 서울시 서대문구 신촌동 134번지 연세대학교 의과대학 예방의학교실, 교수
E-mail: isuh@yumc.yonsei.ac.kr

형(genotype)이다. 따라서 주어진 유전자형으로부터 일배체형을 유추하는 과정이 요구되는데, 이형 접합체(heterozygote)를 갖는 유전자 좌위의 개수가 1개 이하이면 일배체형 짝이 유일하게 결정되지만, 2개 이상이면 $2^{l-1}(l$:이형 접합체 유전자 좌위의 개수)개의 가능한 짝이 존재하여 불확실성이 지수적으로 증가한다. 후자와 같은 경우 분자 생물학적인 분석 방법을 사용하여 일배체형 짝을 유추할 수 있는데 이 방법은 비용이 많이 드는 단점이 있다. 이에 대한 대안으로 통계적 추론을 통해 일배체형을 재구성하는 방법이 널리 사용되고 있다. 대표적인 방법으로는 Clark 알고리즘(Clark, 1990), EM(Expectation-Maximization) 알고리즘(Long 등, 1995), Gibbs sampler 방법(Stephens, 2001), Partition-Ligation 방법(Niu 등, 2002) 등이 있다.

환자-대조군 연관성 분석을 위해 Zhao 등(2000a)은 환자군, 대조군 및 환자군과 대조군을 합친 전체 자료에서 각각 일배체형의 빈도를 추정한 후 우도 함수의 차이를 이용한 우도비 검정법(likelihood ratio test; LRT)을 제안하였다. 환자-대조군 연구 설계를 통해 행해지는 이런 연관성 분석은 소위 인구 집단의 혼합(population admixture) 혹은 인구 집단의 층화(population stratification) 등과 같은 요인이 연구 설계 속에 교락되어 있을 때 왜곡된 분석 결과를 낳을 수 있는 단점이 있다. 한편 부/모/자녀로 구성된 가계 자료에서 널리 사용되는 TDT 방법은 질병이 있는 발단자(proband) 자녀에게 부와 모의 일배체형 짝 중에서 전달되는 일배체형의 빈도와 전달되지 않는 일배체형의 빈도를 비교하는 방법이다. TDT 방법은 가계 내 일배체형의 전달/비전달에 기초하기 때문에 환자-대조군 연관성 분석과 다르게 인구 집단의 혼합이나 층화로부터 자유로운 방법이라고 할 수 있다. Wilson(1997)과 Clayton과 Jones(1999)는 부모의 일배체형이 확실하게 결정된 가계만을 분석에 포함하는 방법을 제안했는데, 이 방법은 부모의 일배체형이 불확실한 가계가 많을 경우 자료의 낭비가 커지게 되므로 비효율적인 방법이라고 할 수 있다. Clayton(1999)은 우도 함수에 근거하여 통계적으로 훌륭한 방법을 제안했는데, 이 방법은 인구 집단의 혼합이나 층화에 민감한 단점을 갖고 있다. Zhao 등(2000b)은 부/모/자녀의 유전자형에 대응하는 가능한 모든 일배체형의 전달/비전달에 대한 조건부 확률로부터, 이에 대응하는 가계의 빈도를 추정하는 한편 이 추정 값으로부터 일배체형의 전달/비전달에 대한 $h \times h(h$:모든 가능한 일배체형의 개수) 2차원 분할표를 만들어 소위 분할표의 주변 동질성(marginal homogeneity)을 검정하는 TDT 통계량을 제안하였다. 이 방법은 Wilson(1997)과 Clayton과 Jones(1999)의 방법과 다르게 부모의 일배체형이 불확실한 가계도 분석에 포함하기 때문에 통계적으로 더 우수한 검정력을 기대할 수 있으며, Clayton(1999)의 방법과 다르게 인구 집단의 혼합이나 층화에 민감하지 않기 때문에 현재까지 개발된 일배체형을 이용한 TDT 방법 중에서는 가장 좋은 특성을 갖는 통계적 방법으로 생각할 수 있다. 김진흠 등(2004)은 일배체형을 이용한 새로운 TDT 통계량을 제안했는데, 이는 Zhao 등(2000b)과 동일한 방법으로 자녀에게 전달되는 일배체형의 전달/비전달에 대한 $h \times h$ 2차원 분할표를 추정한 후, 추정된 분할표의 행 주변합과 이에 대응하는 열 주변합의 차이에 기초한 스코어 통계량이다. Zhao 등(2000b)은 두 주변합 차이의 분산만을 통계량에 고려한 반면, 김진흠 등(2004)은 주변합의 차이들 간에 존재하는 공분산도 함께 고려함으로서 더 많은 정보를 통계량에 포함시켜 검정력을 높이고자 하였다. 김진흠 등(2004)의 모의실험 결과는 김진흠 등(2004)의 스코어

통계량의 검정력이 Zhao 등(2000b)의 그것과 비교하여 서로 비슷하거나 약간 나은 결과를 보여주었다. 여기서 한 가지 주목할 사실은 일배체형의 전달/비전달 2차원 분할표는 단일 유전자 좌위에서 여러 개의 대립 유전자(allele)를 갖는 경우의 분할표와 형태 면으로는 같으나 분할표 내의 각 셀의 자료가 서로 독립이 아니라는 점이 다르다. 독립성이 만족되지 않는 것은 부모의 유전자형 자료로부터 자녀에게 전달된 일배체형을 추론하는 과정에서 불확실성이 존재하여, 부와 모의 일배체형이 자녀에게 전달되는 전달/비전달 여부가 분할표상의 특정한 한 셀에만 기여하는 것이 아니라 여러 셀로 나누어지기 때문이다. 따라서 불행이도 각 셀이 독립이라는 가정 아래서 유도된 'TDT 통계량이 점근적으로 카이제곱 분포를 따른다(Spielamn과 Ewens, 1996)'는 사실을 검정에 이용할 수 없다. 이러한 문제점을 해결하기 위하여 Zhao 등(2000b)과 김진흠 등(2004)은 확률화 검정(randomization test procedure)을 수행하여, 귀무가설 아래서의 경험적 분포를 추정하고 그것으로부터 주어진 분할표의 유의확률을 추정하였다.

ACE는 혈압 상승제로 작용하는 물질인 Angiotensin I을 Angiotensin II로 전환시키는 효소이다. 많은 연구자들이 ACE의 유전적 다형성과 고혈압의 관련성을 규명하기 위한 연구을 진행해왔다(O'Donnell 등, 1998; Tsai 등, 2003; Benjafield 등, 2004).

본 연구에서는 '강화연구'(서일, 2004)의 환자-대조군 자료와 가계 자료를 이용하여 ACE 유전자 내의 4가지 SNP으로 이루어진 일배체형과 고혈압의 관련성을 알아보고자 한다. 2절에서는 강화연구의 배경과 연구 대상자 및 ACE 유전자 내의 SNPs에 대해 설명하고자 한다. 3절에서는 강화자료의 일배체형에 기초한 환자-대조군 연관성 연구와 TDT 연구의 결과를 정리하고자 한다. 또한 이 일배체형을 동정할 수 있는 tag-SNPs에 기초하여 동일한 연구를 하고자 하며, Seltman 등(2003)이 제안한 분기도 분석 방법을 써서 일배체형의 진화과정에서 가깝게 위치하고 질병 발생 위험이 비슷한 클레이드를 찾아내고 일배체형과 고혈압의 연관성을 함께 살펴보고자 한다. 마지막으로 4절에서는 강화자료의 분석 결과 전체에 대해 고찰하고자 한다.

2. 강화연구 자료

2.1. 연구 배경과 연구 대상자

강화연구는 1986년 만 6세에 해당하는 484명의 초등학생들을 대상으로 1997년 17세까지 매년 추적해왔으며, 현재에도 추적 관리되고 있는 코호트 연구로서 한국인의 혈압으로 인한 자연사(natural history)와 혈압 변화의 관련 요인을 찾고자 시도되었다. 강화연구는 코호트 구성원이 중학교에 진학한 1992년 시점에서 715명, 그리고 고등학교로 진학한 1995년 시점에서 784명으로 각각 확대되었으며, 1998년에는 코호트 구성원의 부모에 대한 혈압 및 관련 요인에 대한 조사와 함께 혈액을 채취하였다.

본 연구에 사용된 환자-대조군과 가계 자료는 다음과 같다. 환자군은 15세부터 17세까지 측정한 혈압 중 적어도 한 번 이상 수축기 혈압이 130mmHg 이상이거나 또는 이완기 혈압이 85mmHg인 경우로 정의하였으며, 이 중 혈액이 보관되어 있는 101명(남자 61명, 여자 40명)으로 구성하였다. 대조군은 15세부터 17세까지 측정한 수축기 혈압과 이완기 혈압이

각각 120mmHg와 80mmHg 미만이면서 혈액이 보관되어 있는 구성원 289명 중에서 낮은 혈압 순서대로 176명(남자 74명, 여자 102명)만을 추출하여 구성하였다. JCN-7(Chobanian 등, 2003)의 기준으로 볼 때 환자군의 혈압은 전기 고혈압(pre-hypertension) 이상으로 정의할 수 있고, 대조군은 정상(normal) 혈압군으로 볼 수 있다. 한편 환자군 중에서 부모의 혈액이 모두 있는 40명을 발단자로 하여 부/모/자녀로 이루어진 가계 자료를 구축하였다.

2.2. SNPs

ACE 유전자는 염색체 17q23에 위치하며 총 길이가 26kb 정도이다. Keavney 등(1998)과 Zhu 등(2001)은 ACE 유전자 내에 있는 여러 SNP들을 갖고 연구를 했는데, 본 연구에서는 두 선행 연구에서 사용된 SNP들 중에서 특히 동양인에게 관련성이 예측되는 SNP 4개를 선정하였다. 연구에 포함한 4개 SNP는 각각 Intron 16에 위치한 I/D, Exon 17에 위치한 A2350G, 5′ 프로모터(promotor) 영역에 위치한 A-240T와 C-93T 이다.

3. 연관성 분석

3.1. 환자-대조군 연구

환자-대조군 연구에 포함된 환자군과 대조군의 크기는 각각 101, 176 이었다. 표3.1의 각 열은 환자군, 대조군, 통합군(환자군+대조군)의 자료를 갖고서 EM 알고리즘 방법(Long 등, 1995)으로 추정된 일배체형의 빈도이다. 일배체형에 포함된 SNP들은 순서대로 ACE 유전자 상에 있는 A-240T, C-93T, I/D, A2350G 이다. 대조군이나 환자군에서 공통적으로 TTDG와 ACIA가 각각 30%, 60% 내외로 전체 분포의 90% 정도를 차지하였고, 이 두 일배체형의 4개 모든 SNP가 서로 반대 대립 형질을 갖는 특징을 살펴볼 수 있었다. 그 외 1% 이상의 분포를 갖는 일배체형으로는 대조군의 경우는 TTDA(3%), TTIA(3%), TCDG(1%), ACDG(1%), ACIG(2%), 환자군의 경우는 TTDA(1%), TTIA (2%), TCDA(2%), TCIA(1%), ACIG(1%) 으로 각각 나타났다. 전체적으로 볼 때 90% 정도는 두 집단 간에 큰 차이를 보이지 않지만, 나머지 10% 정도는 두 집단 간에 약간 상이한 모습을 띤다고 할 수 있다.

Zhao 등(2000a)은 일배체형에 기초한 환자-대조군 연관성 검정을 위해 우도비 통계량을 제안하였다. 표3.1의 일배체형에 대한 EM 추정값을 갖고서 계산된 우도비 통계량 값은 19.7 이고, Zhao 등(2000a)이 밝힌 것처럼 유전 형질과 표지 유전자가 서로 연관되어 있지 않다면 우도비 통계량이 근사적으로 자유도가 15인 카이제곱 분포를 따르기 때문에 이 관측된 통계량 값의 유의확률 값(p값)은 0.185이다. 이는 유의수준 5%에서 표지 유전자와 고혈압은 통계적으로 유의하게 연관되어 있지 않다는 것을 의미한다.

3.2. TDT 연구

TDT 연구에서는 환자군에 속한 101명 중에서 40 가계만을 대상으로 하였다. 표3.2는 EM 방법을 써서 부모의 유전자형 자료만 갖고서 일배체형의 빈도를 추정했을 때의 결과와 부모의 유전자형뿐만 아니라 자녀의 유전자형을 포함하여 일배체형의 빈도를 추정했을

표 3.1: 환자-대조군 자료에 대해 EM 방법을 써서 추정된 일배체형의 빈도와 ACE 유전자와 고혈압의 환자-대조군 연관성 검정

	빈도 추정값		
일배체형†	대조군	환자군	통합군
TTDG(1)	0.295	0.340	0.311
TTDA(2)	0.026	0.010	0.020
TTIG(3)	0.002	0.000	0.001
TTIA(4)	0.033	0.015	0.027
TCDG(5)	0.011	0.005	0.009
TCDA(6)	0.000	0.015	0.000
TCIG(7)	0.000	0.005	0.002
TCIA(8)	0.000	0.010	0.003
ATDG(9)	0.000	0.005	0.002
ATDA(10)	0.000	0.000	0.000
ATIG(11)	0.003	0.000	0.002
ATIA(12)	0.000	0.000	0.000
ACDG(13)	0.012	0.000	0.008
ACDA(14)	0.006	0.000	0.004
ACIG(15)	0.015	0.010	0.014
ACIA(16)	0.597	0.600	0.598
$-2\times$로그우도	320.8(df=15)	153.9(df=15)	484.5(df=15)
우도비 통계량		19.7(df=15)	
p값		0.185	

†( )의 숫자는 일배체형의 일련번호를 표시함.

때의 결과를 함께 나열하였다. 전자는 부모의 유전자형에 대응하는 모든 가능한 일배체형 짝을 고려하는 반면에, Rohde와 Fuerst(2001)에 의해서 제안된 후자는 모든 가능한 일배체형 짝 중에서 자녀의 유전자형에 대응되는 짝만을 고려하는 점이 서로 다르다. Rohde와 Fuerst(2001)의 모의실험 결과에 따르면 유전자 좌위의 이질성이 커질수록 후자의 방법이 전자에 비하여 더 우수하고 또한, 자녀의 수가 많아질수록 더 우수한 것으로 나타났다. 하지만 강화연구 자료에서는 그와 같은 두드러진 차이를 찾아 볼 수는 없었다. 두 방법 모두 3.1절의 대조군이나 환자군의 경우와 마찬가지로 TTDG와 ACIA가 전체 분포의 95% 정도를 차지하였고, 그 외 1% 이상을 차지하는 일배체형으로 TTIA(3%)가 있었고, TCIA, ATIA, ACIG는 각각 1% 미만을 차지하는 것으로 나타났다.

Zhao 등(2000b)은 일배체형에 기초한 연쇄 및 연관성 분석을 위해 Spielman과 Ewens (1996)의 TDT 방법을 일배체형으로 확장한 형태의 통계량을 제안하였다. h개 일배체형의

표 3.2: 가계 자료에 대해 EM 방법을 써서 추정된 일배체형의 빈도, 일배체형의 전달/비전달 개체 수와 1,000번의 재표집에 기초한 ACE 유전자와 고혈압의 연쇄 및 연관성에 대한 확률화 검정

	부/모			부/모/자녀		
	빈도	개체 수		빈도	개체 수	
일배체형	추정값	전달	비전달	추정값	전달	비전달
TTDG	0.413	29	37	0.413	29	37
TTDA	0.000	0	0	0.000	0	0
TTIG	0.000	0	0	0.000	0	0
TTIA	0.031	2	3	0.031	2	3
TCDG	0.000	0	0	0.000	0	0
TCDA	0.000	0	0	0.000	0	0
TCIG	0.000	0	0	0.000	0	0
TCIA	0.006	0	1	0.006	0	1
ATDG	0.000	0	0	0.000	0	0
ATDA	0.000	0	0	0.000	0	0
ATIG	0.000	0	0	0.000	0	0
ATIA	0.006	0	1	0.006	0	1
ACDG	0.000	0	0	0.000	0	0
ACDA	0.000	0	0	0.000	0	0
ACIG	0.006	0	1	0.006	0	1
ACIA	0.537	49	37	0.538	49	37
Z		7.363	(p값=0.152)		7.363	(p값=0.152)
S		5.275	(p값=0.229)		5.275	(p값=0.232)

전달/비전달 표를 T라고 하고, T의 한 원소를 $t_{\gamma\delta}$ $(\gamma, \delta = 1, \ldots, h)$라고 하자. 이 때, $t_{\gamma\delta}$은 일배체형 짝 $H_\gamma H_\delta$를 가진 부모 중에서 일배체형 H_γ는 자녀에게 전달하고, 일배체형 H_δ는 전달하지 않은 부모 수를 의미한다. Zhao 등(2000b)의 TDT 통계량 Z는 아래와 같다.

$$Z = \frac{h-1}{h} \sum_{\gamma=1}^{h} \frac{(t_{\gamma\cdot} - t_{\cdot\gamma})^2}{t_{\gamma\cdot} + t_{\cdot\gamma} - 2t_{\gamma\gamma}}.$$

여기서 $t_{\gamma\cdot} = \sum_{\delta=1}^{h} t_{\gamma\delta}$는 일배체형 H_γ를 전달한 부모 수이고, $t_{\cdot\delta} = \sum_{\gamma=1}^{h} t_{\gamma\delta}$는 일배체형 H_δ를 전달하지 않은 부모 수이다. 표3.2의 일배체형에 대한 EM 추정값을 갖고서 Zhao 등(2000b)의 방법대로 일배체형의 전달/비전달 표를 작성하면 가능한 일배체형이 16개 이기 때문에 16×16 표를 얻을 수 있다. 그런데 통계량 Z에서 알 수 있듯이 16×16 표의 각 칸 도수 대신에 각 열과 행의 주변합만으로도 충분히 통계량을 계산할 수 있기 때문에

16×16 표를 제시하는 대신에, 각 일배체형이 전달 또는 비전달된 주변합만을 표3.2에 함께 제시하였다. 예상했던대로 강화연구 자료의 경우는 일배체형 TTDG와 ACIA가 95%를 차지하기 때문에 전달/비전달된 빈도 역시 두 일배체형에 집중된 모습을 볼 수 있다. 이런 현상은 일배체형의 추정 방법에 관계없이 동일하였다. 그런데 통계량 Z가 Spielman과 Ewens(1996)의 그것과 다른 점은 질병 관련 유전자와 표지 유전자가 서로 연쇄되어 있지 않다는 가정 아래서 통계량의 분포가 근사적으로 카이제곱 분포를 따르지 않는다는 것이다. 이런 연유로 관측된 통계량의 유의성을 알아보기 위해 Zhao 등(2000b)은 Zhao 등(1999)이 제안한 확률화 검정 방법을 사용하였다. 한 예를 가지고 이 검정 원리를 설명하면 다음과 같다. 서로 다른 3개의 유전자 좌위가 각각 이대립 형질(diallele) 1과 2를 갖는다고 하자. 한 예로 부의 유전자형은 12/12/11 이고, 모의 유전자형은 22/12/22, 발단자인 자녀의 유전자형은 12/12/12 라고 하자. 모의 유전자형은 오직 두 번째 유전자 좌위에서만 이질적이기 때문에 일배체형 짝은 212와 222로 유일하다. 반면에 부의 유전자형은 첫 번째와 두 번째 유전자 좌위가 모두 이질적이기 때문에 가능한 일배체형 짝이 두 개 존재한다. 그 중 하나는 111과 221 이고, 다른 하나는 121과 211 이다. 따라서 전자의 경우는 부는 111, 모는 222를 전달하는 반면에, 부는 221, 모는 212를 전달하지 못하게 된다. 마찬가지로 후자의 경우는 부는 121, 모는 212를 전달하는 반면에, 부는 211, 모는 222를 전달하지 못하게 된다. 따라서 부와 모가 전달하지 않은 일배체형으로 구성되는 자의 유전자형은 전자의 경우나 후자의 경우 모두 22/12/12 이다. 확률화 검정은 주어진 자료 (부,모,자녀)=(12/12/11, 22/12/22, 12/12/12)와 관측되지 않은 자료 (부,모,자녀)=(12/12/11, 22/12/22, 22/12/12) 중에서 동일한 확률로 무작위 추출하여 표본을 얻고, 그 표본으로부터 통계량을 계산하는 일련의 과정을 반복함으로서 실증적 유의확률을 구하는 방법이다. 일배체형의 추정 방법에 관계없이 통계량 Z의 값은 7.363이고, 상술한 확률화 검정 원리를 따라 1,000번의 재표집을 통해서 계산된 p값은 0.152 였다. 이는 3.1절의 결과처럼 유의수준 5%에서 표지 유전자와 고혈압은 통계적으로 유의하게 연쇄 및 연관되어 있지 않다는 것을 의미한다.

한편 김진흠 등(2004)은 Zhao 등(2000b)의 방법대로 일배체형의 전달/비전달 표 T를 만들고, 그 표에 Stuart(1955)의 방법을 적용한 일명 스코어 검정법을 제안하였다. 김진흠 등(2004)의 TDT 통계량 S는 아래와 같다.

$$S = \Delta' \hat{\Sigma}^{-1} \Delta.$$

여기서 $\Delta = (t_{1\cdot} - t_{\cdot 1}, \ldots, t_{h-1\cdot} - t_{\cdot h-1})'$은 행과 열의 주변합 차이로 이루어진 벡터이고, Δ의 추정 분산-공분산 행렬은

$$\hat{\Sigma} = (\hat{\sigma}_{\gamma\delta}) = \begin{cases} t_{\gamma\cdot} + t_{\cdot\gamma} - 2t_{\gamma\gamma}, & \gamma = \delta \\ -(t_{\gamma\delta} + t_{\delta\gamma}), & \gamma \neq \delta \end{cases}$$

이다. 통계량 S에서 알 수 있듯이 벡터 Δ의 분산뿐만 아니라 공분산도 함께 고려된 점이 통계량 Z와 다른 점이라고 할 수 있다. 그런데 김진흠 등(2004)이 언급한 것처럼 질병 관련 유전자와 표지 유전자가 서로 연쇄되어 있지 않다는 가정 아래서 통계량 S가 근사적으로 카이제곱 분포를 따르지 않기 때문에 Zhao 등(2000b)처럼 확률화 검정을 통해 관측된 통계

량의 실증적 유의확률을 계산하는 방법을 도입하였다. 표3.2에서 볼 수 있듯이 통계량 S의 관측값은 일배체형의 추정 방법에 관계없이 5.275이고, p값은 부모의 정보만 이용한 경우 0.229, 자녀의 정보까지 이용한 경우 0.232로 크게 다르지 않게 나왔다. 이는 통계량 Z를 이용했을 때처럼 유의수준 5%에서 표지 유전자와 고혈압은 통계적으로 유의하게 연쇄 및 연관되어 있지 않다는 것을 의미한다.

3.3. tag-SNPs 연구

본 절에서는 이제까지 연구에 포함했던 SNP 4개 중에 여분(redundant) SNP가 있는지를 알아보고, 여분 SNP를 제거함으로서 연관성 분석에 미치는 영향을 살펴보고자 한다.

표 3.3: 대조군 176명 자료에 대한 D'의 값과 카이제곱 독립성 검정

	A-240T	C-93T	I/D	A2350G
A-240T	-	0.987	0.921	0.860
C-93T	0.000	-	0.869	0.821
I/D	0.000	0.000	-	0.907
A2350G	0.000	0.000	0.000	-

두 SNP 간 연쇄 불평형의 정도를 재기 위한 측도 중에서 널리 사용되는 것으로 Lewontin(1964)에 의해 제안된 D'을 들 수 있는데, 이는 -1과 1사이의 값을 갖기 때문에 서로 다른 유전자 좌위의 LD를 비교하는 데 유용하게 사용된다. 한편 두 유전자 좌위의 대립 형질로 이루어진 2×2 분할표에 잘 알려진 카이제곱 검정법을 적용하여 두 SNP 간 연쇄 불평형의 존재, 다시 말해 두 SNP의 독립성을 검정할 수 있다. 이와 같은 D'의 계산과 독립성 검정은 SAS/Genetics에 있는 Proc Allele(SAS Institute Inc., 2004, pp.17-44)를 통해 쉽게 할 수 있다. 표3.3은 대조군 176명의 자료를 갖고서 구한 4개 SNP간 D' 값(대각선 위)과 카이제곱 독립성 검정의 유의확률 값(대각선 아래)을 나타낸다. 표3.3에서 보여주듯이 4개 SNP 간 어떤 짝도 통계적으로 유의하게 연쇄 불평형 상태에 있음을 알 수 있다. 6개의 가능한 조합 중에서 C-93T와 A2350G의 연쇄 불평형이 제일 낮은 결과를 보여 주었다.

4개 SNP 간 D' 값이 매우 크기 때문에 이들 중에서 여분 SNP를 찾아내어 이 일배체형의 tag-SNP를 정의 할 수 있을 것으로 생각한다. 이를 위해 Clayton(2002)이 제안한 측도 PDE(proportion of diversity explained by a SNP set)를 사용하고자 한다. 회귀 모형의 적합 정도를 알아보기 위한 측도인 결정계수(R^2)처럼 PDE는 일배체형의 다양성을 고려한 SNP 집합으로 설명해 주는 정도를 나타내는 측도이다. PDE의 값은 R^2처럼 tag-SNP 집합에 포함되는 SNP가 많아질수록 증가하는 경향이 있기 때문에 PDE의 증가 폭이 둔화되는 지점에서 tag-SNP 집합을 결정해야 할 것이다(전진 탐색법). 이 방법 외에도 최적의 tag-SNP 집합을 찾기 위한 여러 방법들이 SAS/Genetics에 있는 Proc Htsnp(SAS Institute Inc., 2004, pp.125-138)에 소개되어 있다. 대조군 176명의 자료에 대해 전진탐색법을 적용해 본 결과 최적의 tag-SNP 집합은 C-93T와 A2350G로 나타났으며, 이 때 PDE의 값은 0.992로 나타

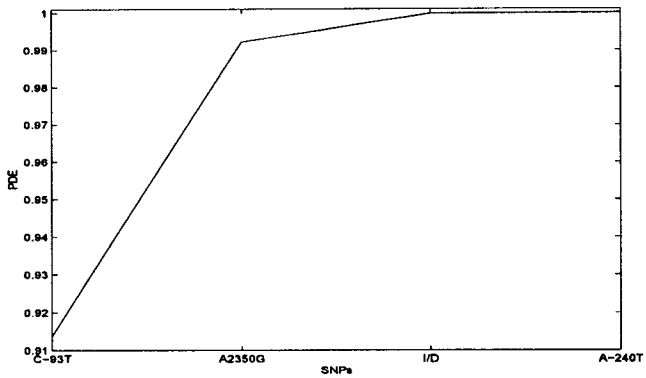

그림 3.1: tag-SNP 집합에 SNP의 추가에 따른 PDE의 변화

났다. 그림3.1에서 볼 수 있듯이 SNP C-93T 만을 tag-SNP 집합에 포함했을 때는 PDE가 0.914로 A2350G를 추가함으로서 PDE가 0.078이 증가한 반면에, C-93T와 A2350G로 이루어진 tag-SNP 집합에 SNP I/D를 추가함으로서 0.008, C-93T와 A2350G, I/D로 이루어진 집합에 A-240T를 추가함으로서 0.0002가 각각 증가하여 그 폭이 매우 작음을 알 수 있었다. 한편 두 SNP C-93T와 A2350G 만을 갖고서 3.1절과 3.2절에서 소개한 연관성 분석을 각각 수행하였다. 우도비 검정의 결과는 통계량 값(p값)이 5.183(0.159)으로 4개의 SNP를 갖고 연관성 분석을 했을 때 보다는 p값이 작게 나와 tag-SNPs 만을 갖고 분석했을 때가 상대적으로 약간 더 유의한 결과를 보여주었지만, 이 때에도 여전히 p값이 충분히 작지 않아 tag-SNPs과 고혈압이 통계적으로 유의하게 연관되어 있다고는 할 수 없다. 1,000번의 재표집을 통한 TDT 결과를 살펴보면, Zhao 등(2000b)의 통계량(p값) 값은 4.314(0.163), 김진흠 등(2004)의 통계량(p값)은 3.318(0.332)로 각각 나타났다. 이 때에도 역시 tag-SNPs과 고혈압은 통계적으로 유의하게 연쇄 및 연관되어 있지 않다는 것을 의미한다.

3.4. 분기도 연구

본 절에서는 Seltman 등(2003)이 개발한 EHAP(Evolutionary-Based Haplotype Analysis Package) 프로그램을 이용하여 분기도 분석을 수행하여 일배체형의 진화 과정에서 가깝게 위치하고 질병 발생 위험이 비슷한 클레이드(clade)를 찾아내고 일배체형과 고혈압의 연관성을 살펴보고자 한다. 그림3.2과 그림3.3에 표시된 숫자는 일배체형을 의미하며 그것에 대응하는 SNP 표현은 표3.1의 1열에서 찾아볼 수 있다. 한 선으로 연결된 두 일배체형들은 오직 한 유전자 좌위에서만 서로 다른 대립 형질을 갖는다.

그림3.2의 왼쪽 그림은 환자-대조군 자료에서 0.0001 이상의 빈도를 갖는 13개의 일배체형에 대한 분기도이다. 독립 변수로 13개 일배체형을 갖는 로지스틱 회귀모형을 고려하고 우도비 검정을 통해 클레이드를 조사해본 결과, 그림3.2의 오른쪽 그림과 같이 모든 대립 형질이 wild-type인 ACIA(16)으로만 구성된 클레이드(clade a)[59.8%], 일배체형 1과

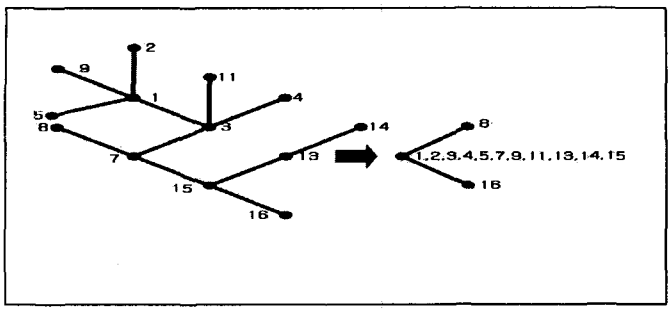

그림 3.2: 환자-대조군 자료에 대한 분기도(왼쪽)와 우도비 검정에 의해 추정된 클레이드(오른쪽)

비교하여 첫번째 유전자 좌위의 대립 형질만 다른 TCIA(8)으로만 구성된 클레이드(clade b)[0.3%], 그외 나머지 11개의 일배체형으로 구성된 클레이드(clade c)[39.9%]로 분리됨을 알 수 있었다. 어떤 개체가 'clade a', 'clade b', 혹은 'clade c'에 속하는 일배체형을 갖고 있을 때, 가변수 벡터를 각각 $(d_1, d_2) = (0,0), (1,0), (0,1)$라고 하자. 이 때, 추정된 회귀식은 $logit = -0.560 + 21.2 * d_1 - 0.040 * d_2$ 이며, 전체 분포의 99.7%를 차지하는 'clade a'와 'clade c'는 고혈압 발생 위험에 있어 차이가 없지만 전체 분포의 0.3%를 차지하는 'clade b'는 'clade a'에 비해 고혈압 발생 위험이 매우 높은 것으로 나타났다(p값=0.000007).

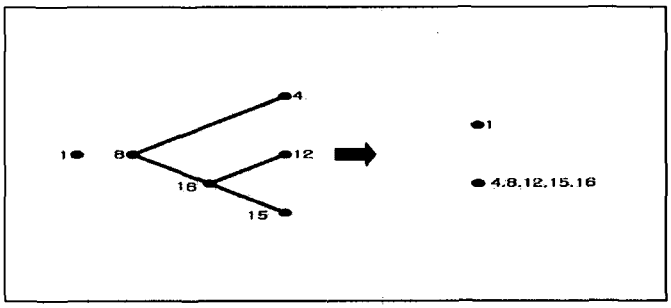

그림 3.3: 가계 자료에 대한 분기도(왼쪽)와 우도비 검정에 의해 추정된 클레이드(오른쪽)

그림3.3의 왼쪽 그림은 가계 자료에서 0.0001 이상의 빈도를 갖는 6개의 일배체형에 대한 분기도이고, 오른쪽 그림은 이를 독립 변수로 하는 로지스틱 회귀모형을 고려하고 스코어 검정을 통해 얻어진 클레이드이다. 이들은 모든 대립 형질이 mutant-type인 TTDG(1)로만 구성된 클레이드(clade a)[41.3%]와 그외 나머지 5개의 일배체형으로 구성된 클레이드(clade b)[58.7%]로 분리되었다. 발단자인 자녀에게 TTDG는 전달되고 TTIA, TCIA, ACIA는 전달되지 않는 빈도가 각각 1, 1, 15 이기 때문에, 결국 'clade a'는 전달되고 'clade b'는 전달되지 않는 빈도가 17이 된다. 또한 ACIA는 발단자에게 전달되고 TTDG는 전달되지 않는

빈도가 25 이므로 'clade b'는 전달되고 'clade a'는 전달되지 않는 빈도가 25로 그 반대의 경우보다 약간 높게 나타났다 [표3.4].

표 3.4: 가계 자료의 분기도 분석에 기초한 일배체형의 전달/비전달 개체 수

		비전달된 일배체형		
		clade a	clade b	합계
전달된	clade a	12	17	29
일배체형	clade b	25	26	51
	합계	37	43	80

4. 결론 및 고찰

이제까지 강화연구 자료를 갖고 ACE 유전자와 고혈압의 연관성을 살펴보았다. 환자군 101명과 대조군 176명에 기초한 환자-대조군 연관성 연구에서는 4개의 SNP -A-240T, C-93T, I/D, A2350G-로 이루어진 일배체형과 고혈압이 서로 연관되어 있지 않은 것으로 나타났으며, PDE 측도에 기초한 전진 탐색법에 의해 결정된 tag-SNPs -C-93T와 A2350G-과 고혈압의 연관성 검정에서도 통계적으로 유의하지 않다는 결론에 도달했다. 동일한 자료에 대한 분기도 분석에서 발견된 특징은 전체 분포의 0.3%를 차지하는 일배체형 TCIA가 고혈압에 미치는 영향이 매우 크다는 점이다. 한편 40 가계 자료에 기초한 TDT 연구에서도 동일하게 4개의 SNP과 C-93T와 A2350G으로만 이루어진 tag-SNPs을 갖고서 Zhao 등(2000b)과 김진흠 등(2004)의 TDT 통계량을 써서 일배체형과 고혈압의 연쇄 및 연관성을 검정했는데 통계적으로 유의한 결과를 발견하지 못했다. 분기도 분석에서는 모든 유전자 좌위의 대립 형질이 서로 반대인 두 일배체형 TTDG와 ACIA의 전달/비전달 개체 수를 비교해보면, ACIA의 전달이 TTDG 보다 상대적으로 많은 것으로 나타났는데 이 결과는 환자-대조군 연구에서 TTDG의 빈도가 대조군에서 보다 오히려 환자군에서 많았던 결과와 서로 상반된다. 본 연구의 제한점은 연구 대상자 수가 이들의 연관성을 밝히기에 충분하지 않았으며, 고혈압을 정의하는데 있어 연구 대상자 수가 적어 그 기준을 낮춤으로서 이들 연관성의 강도가 약화되었을 가능성도 배제할 수는 없다. 추후 충분한 연구 대상자를 확보하고, ACE 유전자 내의 한국인의 특이 SNP을 이용하여 연구를 할 필요성이 있다고 생각한다.

참고문헌

김진흠, 강대룡, 이윤경, 신선미, 서일, 남정모 (2004). 일배체형에 기초한 연쇄분석의 통계학적 알고리즘의 연구, <예방의학회지>, **37**, 366-372.

서일 (2004). 혈압의 장기 변화와 고혈압 발생에 대한 유전역학적 연구, 한국학술진흥재단.

Benjafield, A. V., Wang, W. Y., and Morris, B. J. (2004). No association of angiotensin-converting enzyme 2 gene (ACE2) polymorphisms with essential hypertension, *American Journal of Hypertension*, **17**, 624-628.

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L., Jones, D. W., Materson, B. J., Oparil, S., Wright, J. T., Roccella, E. J., and the National High Blood Pressure Education Program Coordinating Committee. (2003). The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure(JCN-7), *Journal of the American Medical Association*, **289**, 2560-2572.

Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid population, *Molecular Biology and Evolution*, **7**, 111-122.

Clayton, D. (1999). A generalization of the transmission/disequilibrium test(TDT) for uncertain haplotype transmission, *American Journal of Human Genetics*, **65**, 1170-1177.

Clayton, D. (2002). Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci, http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP/htsnp.pdf.

Clayton, D. and Jones, H. (1999). Transmission/disequilibrium tests for extended marker haplotypes, *American Journal of Human Genetics*, **65**, 1161-1169.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome, *Science*, **296**, 2225-2229.

Jorde, L. B. (1995). Linkage disequilibrium as a genetic mapping tool, *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.

Keavney, B., McKenzie, C. A., Connell, J. M., Julier, C., Ratcliffe, P. J., Sobel, E., Lathrop, M., and Farrall, M. (1998). Measured haplotype analysis of the angiotensin-I converting enzyme gene, *Human Molecular Genetics*, **7**, 1745-1751.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models, *Genetics*, **49**, 49-67.

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

Niu, T., Qin, Z. S., Xu, X., and Liu, J. S. (2002). Bayesian haplotype inference for multiple linked single nucleotide polymorphisms, *American Journal of Human Genetics*, **70**, 157-169.

O'Donnell, C. J., Lindpaintner, K., Larson, M. G., Rao, V. S., Ordovas, J. M., Schaefer, E. J., Myers, R. H., and Levy, D. (1998). Evidence for association and genetic linkage of the angiotensin-converting enzyme locus with hypertension and blood pressure in men but not women in the Framingham Heart Study, *Circulation*, **97**, 1766-1772.

Rohde, K. and Fuerst, R. (2001). Haplotying and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information, *Human Mutation*, **17**, 289-295.

SAS Institute Inc. (2004). *SAS/Genetics 9.1 User's Guide*, SAS Institute Inc., Cary.

Seltman, H., Roeder, K., and Devlin, B. (2003). Evolutionary-based association analysis using haplotype data, *Genetic Epidemiology*, **25**, 48-58.

Spielman, R. S. and Ewens, W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association, *American Journal of Human Genetics*, **59**, 983-989.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data, *American Journal of Human Genetics*, **68**, 978-989.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification, *Biometrika*, **42**, 412-416.

The International HapMap Consortium. (2003). The international HapMap project, *Nature*, **426**, 789-796.

Tsai, C. T., Fallin, D., Chiang, F. T., Hwang, J. J., Lai, L. P., Hsu, K. L., Tseng, C. D., Liau, C. S., and Tseng, Y. Z. (2003). Angiotensinogen gene haplotype and hypertension: interaction with ACE gene I allele, *Hypertension*, **41**, 9-15.

Wilson, S. R. (1997). On extending the transmission/disequlibrium test(TDT), *Annals of Human Genetics*, **61**, 151-161.

Zhao, H., Kathleen, R. M., and Kenneth, K. K. (1999). On a randomization procedure in linkage analysis, *American Journal of Human Genetics*, **65**, 1449-1456.

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

Zhao, H., Zhang, S., Merikangas, K. R., Trixler, M., Wildenaauer, D. B., Sun, F., and Kidd, K. K. (2000b). Transmission/disequilibrium tests using multiple tightly linked markers, *American Journal of Human Genetics*, **67**, 936-946.

Zhu, X., Bouzekri, N., Southam, L., Cooper, R. S., Adeyemo, A., McKenzie, C. A., Luke, A., Chen, G., Elston, R. C., and Ward, R. (2001). Linkage and association analysis of angiotensin I-converting enzyme (ACE)-gene polymorphisms with ACE concentration and blood pressure, *American Journal of Human Genetics*, **68**, 1139-1148.

[2005년 1월 접수, 2005년 3월 채택]

Haplotype-Based Association and Linkage Analysis of Angiotensin-I Converting Enzyme(ACE) Gene with a Hypertension*

Jinheum Kim[1] Chung Mo Nam[2] Dae Ryong Kang [3] Il Suh[4]

ABSTRACT

In this study we investigate the association between the haplotype block of 4 SNPs in ACE genes and hypertension with a case-control dataset of size of 277 and 40 families data collected from Kangwha studies. To this end we perform a haplotype-based case-control association study and a haplotype-based TDT study. We do the same analysis with tag-SNPs that can identify the haplotype block. Through a cladogram analysis we make the evolution-tree of haplotypes and then classify the haplotypes into a few clades by collecting haplotypes exposed to the disease to the same extent. We also discuss the association between these clades and hypertension.

Keywords: ACE genes; Case-control study; Families data; Haplotypes; TDT

* This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health and Welfare(03-PJ1-PG3-21000-0015, Chung Mo Nam).

1) Associate Professor, Department of Applied Statistics, University of Suwon, Gyeonggi-Do, 445-743, South Korea
E-mail:jinhkim@suwon.ac.kr

2) Associate Professor, Department of Preventive Medicine and Public Health, Yonsei University College of Medicine, C.P.O. Box 8044, Seoul, 120-749, South Korea
E-mail: cmnam@yumc.yonsei.ac.kr

3) Research Fellow, Department of Preventive Medicine and Public Health, Yonsei University College of Medicine, C.P.O. Box 8044, Seoul, 120-749, South Korea
E-mail: cosakang@yumc.yonsei.ac.kr

4) Professor, Department of Preventive Medicine and Public Health, Yonsei University College of Medicine, C.P.O. Box 8044, Seoul, 120-749, South Korea
E-mail: isuh@yumc.yonsei.ac.kr