예방의학회지 제37권 제4호(2004년 11월)
J Prev Med Public Health 2004;37(4):366-372

일배체형에 기초한 연쇄분석의 통계학적 알고리즘 연구

김진흠, 강대룡[1], 이윤경[2], 신선미[2], 서 일[1], 남정모[1]

수원대학교 통계정보학과, 연세대학교 의과대학 예방의학교실[1],
연세대학교 대학원 보건학과[2]

Statistical Algorithm in Genetic Linkage Based on Haplotypes

Jinheum Kim, Dae Ryong Kang[1], Yun Kyung Lee[2], Sun Mi Shin[2], Il Suh [1], Chung Mo Nam[1]

Department of Applied Statistics, University of Suwon,
Department of Preventive Medicine and Public Health, Yonsei University College of Medicine[1],
Graduate School of Public Health, Yonsei University[2]

Objectives : This study was conducted to propose a new transmission/disequilibrium test(TDT) to test the linkage between genetic markers and disease-susceptibility genes based on haplotypes. Simulation studies were performed to compare the proposed method with that of Zhao et al. in terms of type I error probability and powers.

Methods : We estimated the haplotype frequencies using the expectation-maximization(EM) algorithm with parents' genotypes taken from a trio dataset, and then constructed a two-way contingency table containing estimated frequencies to all possible pairs of parents' haplotypes. We proposed a score test based on differences between column marginals and their corresponding row marginals. The test also involved a covariance structure of marginal differences and their variances. In simulation, we considered a coalescent model with three genetic markers of biallele to investigate the performance of the proposed test under six different configurations.

Results : The haplotype-based TDT statistics, our test and Zhao et al.'s test satisfied a type I error probability, but the TDT test based on single locus showed a conservative trend. As expected, the tests based on haplotypes also had better powers than those based on single locus. Our test and that of Zhao et al. were comparable in powers.

Conclusion : We proposed a TDT statistic based on haplotypes and showed through simulations that our test was more powerful than the single locus-based test. We will extend our method to multiplex data with affected and/or unaffected sibling(s) or simplex data having only one parent's genotype.

J Prev Med Public Health 2004;37(4):366-372

Key Words: Transmission/disequilibrium test, Linkage, Haplotypes, Score test, Simulation

서 론

질병 관련 유전자의 위치를 탐색하는 유전역학 방법으로 유전자좌위(loci)의 재조합비율(recombinant fraction)을 이용한 연쇄분석(linkage analysis)과 대립유전자의 불평형(linkage disequilibrium: LD) 정도를 이용하는 연관성분석(allelic association analysis)이 대표적이며, 이들 방법들은 서로 상호 보완적인 관계에 있다. 연관성분석은 환자-대조군 연구설계를 통해 이루어지는데 인구집단의 혼합(population admixture) 또는 인구집단의 층화(population stratification)가 있는 경우에 방법론적으로 심각한 문제가 발생할 수 있는 단점을 갖고 있지만, 인구집단의 오랜 역사를 거쳐 현재 존재하는 대립유전자의 불평형을 비교하므로 좁은 영역의 위치탐색(fine mapping)이 가능한 장점이 있다. 연쇄분석은 발단자(proband)를 중심으로 이루어진 여러 세대의 가계자료에 대해 특정한 유전모형을 가정하고, 이를 통해 표지유전자(genetic marker)와 질병의 표현형(phenotype)으로부터 재조합비율과 로드값(lod score)을 추정하는 고전적인 방법이 현재까지 가장 많이 사용되고 있다. 그러나 이 방법은 다인성 질환(multifactorial disease) 또는 복합성 질환(complex disease)과 같이 발단자의 연령이 높은 경우, 여러 세대의 자료를 수집하기 어려울 뿐만 아니라 특정한 유전모형을 가정하기 어려운 문제점을 갖고 있다. 대안으로 현재 TDT (transmission/disequilibrium test)와 이의 확장된 방법이 널리 사용되고 있다 [1]. TDT 통계량은 연쇄분석과 연관성분석의 의미를 동시에 가지며, 표지유전자와 질병관련 유전자 사이에 LD가 존재하는 경우에 통계적인 검정력이 매우 우수하고 인구집단의 혼합이나 층화에 대해 민감하지 않는 좋은 특성을 가지고 있다 [2].

최근 염기서열 분석기술의 빠른 발전, 국제간 SNP(single nucleotide polymorphism) 협력연구 [3]와 HapMAP 프로젝트의 연구결과 [4,5]들은 유전자의 위치를 탐색하는 방법에 큰 변화를 주고 있다. 즉, 유전자의 위치를 탐색하는데 있어, 하나의 표지유전자 또는 SNP을 이용하는 방법으로부터 유전체 상에 가깝게 위치한 SNP들의 일배

접수 : 2004년 6월 23일, 채택 : 2004년 8월 18일
본 연구는 보건복지부 보건의료기술진흥사업의 지원에 의하여 이루어진 것임.(03-PJ1-PG3-21000-0015)
책임저자 : 남정모(서울 서대문구 신촌동 134, 전화 : 02-361-5387, 팩스 : 02-392-8133, E-mail : cmnam@yumc.yonsei.ac.kr)

체형(haplotype)을 이용하여 탐색하는 방향으로 변화되고 있다. 이러한 연구방향의 변화는 단일 유전자좌위의 정보를 이용한 연쇄분석 또는 연관성분석보다 일배체형의 정보를 이용한 분석의 통계적 검정력이 높기 때문이다 [6,7]. SNP들은 유전체 상에서 랜덤하게 분포하기 보다는 블록(block)의 형태로 존재하는데, 일배체형 블록을 대표할 수 있는 적은 수의 tag SNP들을 사용하여 유전자의 위치를 탐색하는 방법이 현재 당면하고 있는 유전체 연구에서의 많은 문제점을 해결할 수 있다 [8].

일배체형에 기초한 연쇄분석과 연관성분석을 위해서는 먼저 각 개인의 일배체형을 알아야하는데, 값 비싼 분자생물학적 분석(molecular haplotyping)이 아니고는 통계적인 추론으로 일배체형을 재구성(haplotype reconstruction)할 수밖에 없으며 현재까지 이 방법을 보편적으로 사용하고 있다. 일배체형을 재구성할 때 사용되는 통계적 방법으로는 Clark 알고리즘 [9], EM(Expectation-Maximization) 알고리즘 [10], Gibbs sampler 방법 [11], Partion - Ligation 방법 [12] 등이 있으며 서로 장·단점이 있다.

최근 연관성분석과 연쇄분석에서 일배체형을 이용한 통계적 방법들이 개발되고 있다. 환자-대조군의 연관성분석에서는 환자군, 대조군, 환자군과 대조군을 합친 전체자료에서 각각 일배체형의 빈도를 추정한 후 우도함수(likelihood function)의 차이를 이용한 우도비 검정(likelihood ratio test; LRT) 방법을 사용하고 있다 [13]. 연쇄분석에서는 부/모/자로 구성된 trio 자료를 이용한 TDT 방법을 사용하고 있는데, 이는 기본적으로 질병이 있는 발단자 자녀에게 부와 모의 일배체형 짝 중에서 어떤 일배체형이 전달되고 어떤 일배체형이 전달되지 않느냐에 대한 빈도를 추정하여 비교하는 방법이다. 이러한 접근의 방법론적 문제는 일배체형의 전달/비전달(transmission/non-transmission) 빈도를 주어진 유전자형(genotype) 자료로부터 추정할 때 불확실성(uncertainty)이 존재한다는 점이다. 즉, k개의 유전자좌위에서, c개의 이형접합성(heterozygote)이 있는 대상은 $2^{(c)}$개의 가능한 일배체형 짝이 있다. 따라서 trio 자료에서 일배체형의 전달/비전달은 상당히 복잡한 구조를 가진다. Wilson [14]과 Clayton과 Jones [15]는 일배체형의 전달/비전달이 불확실한 가계는 분석에서 제외하는 방법을 제안하였으나, 자료의 낭비가 커지게 되므로 비효율적이다. Clayton [16]은 우도함수에 근거하여 통계적으로 좋은 방법을 제안하였으나, 이 방법은 인구집단의 혼합 또는 층화에 민감한 문제점을 가지고 있다. 한편 Zhao 등 [17]은 부/모/자의 유전자형 자료에서 모든 가능한 일배체형의 전달/비전달에 대한 조건부 확률로부터, 이에 대응하는 가계의 빈도를 추정하고 또한 이 추정 값으로부터 일배체형의 전달/비전달에 대한 2차원 분할표를 만들어 분할표의 주변동질성 검정(marginal homogeneity)을 이용한 방법을 제안하였다. 이 방법은 Wilson[14]과 Clayton과 Jones [15]의 방법과 다르게 불확실성을 갖는 가계도 분석에 포함하기 때문에 더 우수한 통계적 검정력을 기대할 수 있으며, Clayton [16]의 방법과 다르게 인구집단의 혼합 또는 층화에 민감하지 않기 때문에 현재까지 개발된 일배체형을 이용한 TDT 방법 중 가장 좋은 특성을 갖는 통계적 방법으로 생각할 수 있다.

본 연구는 일배체형을 이용한 새로운 TDT 통계량을 제안하고, 제안한 새로운 방법을 Zhao 등 [17]의 일배체형을 이용한 TDT 통계량과 시뮬레이션을 통하여 그 특성을 비교하였으며, 이 두 방법을 기존의 단일 유전자좌위의 TDT 방법과 서로 비교하였다. 또한 환자-대조군에 대한 단일 유전자좌위의 유전자형에 기초한 연관성분석과 일배체형에 기초한 연관성분석을 동시에 수행함으로서 연쇄분석과 연관성분석의 검정력을 비교하였다.

연구 방법

본 연구는 다음 두 가지 단계로 진행되었다. 첫 번째 단계는 부/모/자로 구성된 trio 자료에서 일배체형에 기초한 새로운 TDT 통계량을 제안하는 단계이며, 두 번째 단계는 새롭게 제안한 방법과 기존의 방법들을 시뮬레이션을 통하여 제 1종의 오류와 검정력을 비교하는 단계이다.

1. 일배체형에 기초한 새로운 TDT 통계량

일반적으로 일배체형에 기초한 TDT 통계량은 부모의 일배체를 자녀에게 전달할 때 어떤 일배체를 전달하는지에 대한 전달/비전달의 불확실성이 존재한다. 예를 들어 유전자좌위가 3개이고 각각의 유전자좌위는 2개의 대립유전자 1과 2를 갖는 경우를 생각해보자. 부의 유전자형이 12/12/11, 모의 유전자형이 22/12/22, 자의 유전자형이 12/12/12인 trio에서, 부의 가능한 일배체형의 짝은 {111/221} 또는 {121/211}의 두 가지 형태가 있고, 모의 가능한 일배체형의 짝은 {212/222} 한가지뿐이다. 이러한 경우, 자녀의 유전자형을 고려할 때, 만약 부의 일배체형 짝이 {111/221}이면 부는 자에게 {111}로 구성된 일배체형, 모는 {222}로 구성된 일배체형을 전달하여야 한다. 또한 부의 일배체형 짝이 {121/211}이면 부는 자에게 {121}로 구성된 일배체형을, 모는 {212}로 구성된 일배체형을 전달하여야 하므로 이러한 trio는 일배체형의 전달/비전달에 불확실성이 존재한다[1].

Zhao 등 [17]은 먼저 부모의 유전자형 자료로부터 먼저 일배체형의 비율을 EM 알고리즘을 사용하여 추정하고 (h개의 일배체형이 있다고 가정함), 부/모/자의 유전자형 패턴이 동일한 가계의 집합에서 모든 가능한 일배체형 짝의 집합을 구성한 후, 조건부 확률을 이용하여 부와 모의 특정한 일배체형의 전달/비전달에 대한 결합확률(joint probability) 값을 추정하여 이에 대응하는 가계의 빈도를 추정하고, 또한 이 추정 값으로부터 모든 가능한 일배체형 짝의 전달/비전달에 대한 $h \times h$분할표를 구성하여 주변합(marginal sum)을 이용한 주변동질성 검정을 위한 방법을 제안하였다. 이 방법은 분할표가 구성되면, 여러 개(>2)의 대립유전자를 갖는 단일 유전자좌위에서의 TDT 통계량 [18]과 동일한

1) 이 경우 부모와 자녀간의 해당 일배체형 블록내에 재조합이 없다고 가정함.

방법이다. 그러나 일배체형의 전달/비전달 $h \times h$분할표에서 각 셀이 서로 독립이 아니기 때문에 Spielman과 Ewens [18]처럼 TDT 통계량의 점근적인 분포를 이용하여 검정할 수 없고, Zhao 등 [17]은 확률화 검정(randomi-zation test)을 적용하여 경험적(empirical) 유의확률을 추정하는 방법을 아울러 제안하였다.

본 연구에서 제안하는 방법은 두 단계로 나누어 설명할 수 있다. 첫 번째는 일배체형의 전달/비전달에 대한 $h \times h$분할표를 추정하는 방법으로서 본 연구에서는 Zhao 등[17]의 방법과 동일한 방법을 사용하였다. 두 번째는 통계량의 형태와 이론적 근거로서, Zhao 등 [17]의 방법은 Spielman과 Ewens [18]의 통계량을 사용하였으나, 본 연구에서는 Stuart [19]의 통계량을 사용하였다. 위의 두 통계량은 첫 번째 단계에서 구축된 전달/비전달 분할표의 행(row) 주변합과 대응하는 열(column) 주변합 차이를 이용하는 측면은 같다. 그러나 전자는 주변합 차이의 분산만을 이용하였고, 후자는 주변합 차이의 공분산을 추정하고 이를 통계량에 포함함으로서 더 많은 정보를 활용하였다. 먼저 $\hat{T}_{i\cdot}$를 전달된 i번째 일배체형의 행 주변합 $\hat{T}_{\cdot j}$를 전달되지 않은 j번째 일배체형의 열 주변합, $\hat{T}_{ij}$를 일배체형 짝 중에서 i번째 일배체형은 전달되고 j번째 일배체형은 전달되지 않은 빈도수로 각각 정의할 때, 스코아 통계량은 다음과 같다.

$T = \mathbf{a}'\Sigma^{-1}\mathbf{a}$ 여기서,

$\mathbf{a} = (\hat{T}_{1\cdot} - \hat{T}_{\cdot 1}, \ldots, \hat{T}_{(h-1)\cdot} - \hat{T}_{\cdot (h-1)})'$

이며, 분산-공분산 행렬은

$$\Sigma = (\sigma_{ij}) = \begin{cases} \hat{T}_{i\cdot} + \hat{T}_{\cdot i} - 2\hat{T}_{ii}, & i = j \\ -(\hat{T}_{ij} + \hat{T}_{ji}), & i \neq j \end{cases}$$

이다.

통계량 T_S는 Zhao 등 [17]의 통계량과 마찬가지로 분할표의 각 셀들이 서로 독립이 아니기 때문에 본 연구에서도 확률화 검정을 통하여 제안한 스코아 통계량의 경험적 유의확률을 추정하였다. 확률화 검정은 자녀에게 전달된 부/모의 일배체형 짝과 전달되지 않은 일배체형 짝 중 동일한 확률로 랜덤하게 하나의 일배체형 짝을 추출한 후, 이를 자녀에게 전달함으로서 자녀의 유전자형을 새로 생성하여 경험적 유의확률을 추정하는 방법이다.

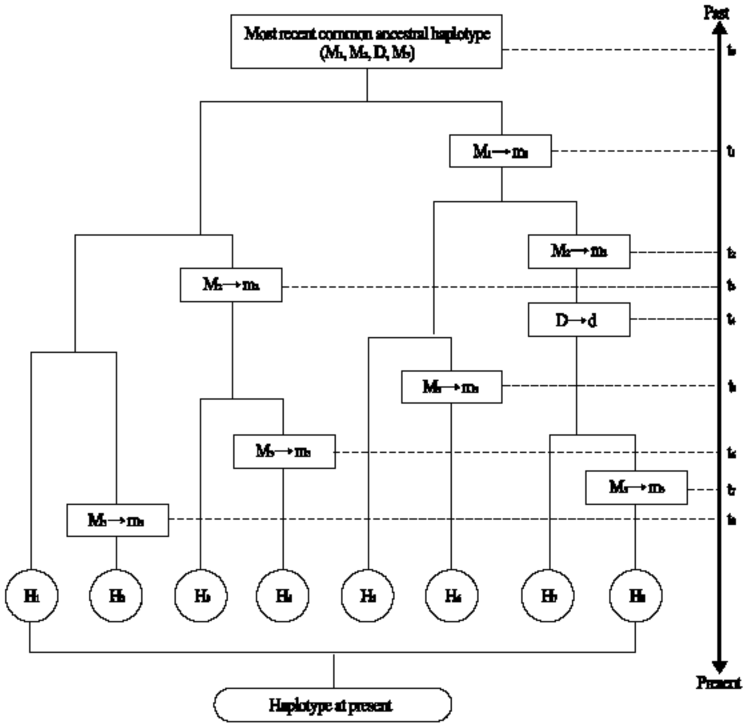

Figure 1. Genealogy of four mutations at four loci used in simulation study.

2. 시뮬레이션 모형

본 연구의 시뮬레이션은 일배체형 블록에 3개의 표지유전자좌위, 그리고 각각의 표지유전자좌위에는 2개의 대립유전자로 구성된 경우로만 국한하였다. 즉, 3개의 표지유전자좌위의 대립유전자들을 각각 $\{M_1, m_1\}$, $\{M_2, m_2\}$, $\{M_3, m_3\}$로 정의하고, 관측되지 않았지만 질병발생과 관련된 유전자좌위의 대립유전자를 $\{D, d\}$로 정의하였다. 연구대상 유전자형 자료를 주어진 조건하에서 발생하기 위하여 Figure. 1과 같은 모형을 고려하였다. 이 모형에서 질병발생과 관련된 대립유전자 d를 가진 대상은 첫 번째와 두 번째 표지유전자좌위의 대립유전자가 각각 변이된 m_1과 m_2만을 가지고 있으므로 이들 표지유전자와 질병관련 유전자는 LD가 높으나[2], 세 번째 표지유전자좌위의 대립유전자는 M_3와 m_3로 다양하므로 LD가 낮음을 알 수 있다. 또한 8가지 가능한 일배체형 중 7번째와 8번째 일배체형 (m_1, m_2, M_3)와 (m_1, m_2, m_3)이 질병발생과 관련된 대립유전자 d를 포함하고 있음을 알 수 있다.

연구대상의 표현형인 질병유무에 대한 자료는 다음과 같은 사실에 근거하여 생성하였다. 일반적으로 다인성질환은 질병발생과 관련된 대립유전자 d를 가지고 있더라도 질병이 100% 발생하지 않는 침투율(penetrance)을 보이며, 또한 대립유전자 d를 가지고 있지 않더라도 질병발생 위험이 존재한다. 따라서 본 연구의 시뮬레이션에서는 질병발생과 관련된 대립유전자 d를 가지고 있지 않는 대상의 질병발생 가능성은 주어진 기저위험(background risk) 비율에 따라 확률적으로 발생하고, 대립유전자 d가 있는 대상의 질병발생 가능성은 기저위험에 질병발생 상대위험(relative risk: RR)을 곱한 크기에 따라 확률적으로 발생하였다. 단, 질병발생에 대한 유전모형(inheritance of mode)은 대립유전자 d에 대한 우성모형(dominant model)을 가정하였다. 따라서 RR이 1이면, 조사된 표지유전자는 질병발생과 관련된 참(true) 유전자좌위와 연쇄되지 않았으며 또한 연관성이 없는 것으로 해석할 수 있고, 반면 RR이 1보다 커지면 연쇄되어 있거나 연관성이 있는 것으로 해석할 수 있다. 시뮬레이션

2) LD 중 일반적으로 많이 사용하는 D'이 1임.

Table 1. Types of haplotype frequencies and background disease risk in each population

Configuration	Population	Background disease risk	Haplotype frequency* (H_1, H_2, H_3, H_4, H_5, H_6, H_7, H_8)
I	1	0.1	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
	2	0.1	(0.490, 0.000, 0.210, 0.000, 0.210, 0.000, 0.090, 0.000)
II	1	0.2	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
	2	0.1	(0.490, 0.000, 0.210, 0.000, 0.210, 0.000, 0.090, 0.000)
III	1	0.1	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
	2	0.1	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
IV	1	0.2	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
	2	0.1	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
V	1	0.1	(0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125)
	2	0.1	(0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125)
VI	1	0.2	(0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125)
	2	0.1	(0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125)

* $H_1=(M_1, M_2, M_3)$, $H_2=(M_1, M_2, m_3)$, $H_3=(M_1, m_2, M_3)$, $H_4=(M_1, m_2, m_3)$, $H_5=(m_1, M_2, M_3)$, $H_6=(m_1, M_2, m_3)$, $H_7=(m_1, m_2, M_3)$, $H_8=(m_1, m_2, m_3)$

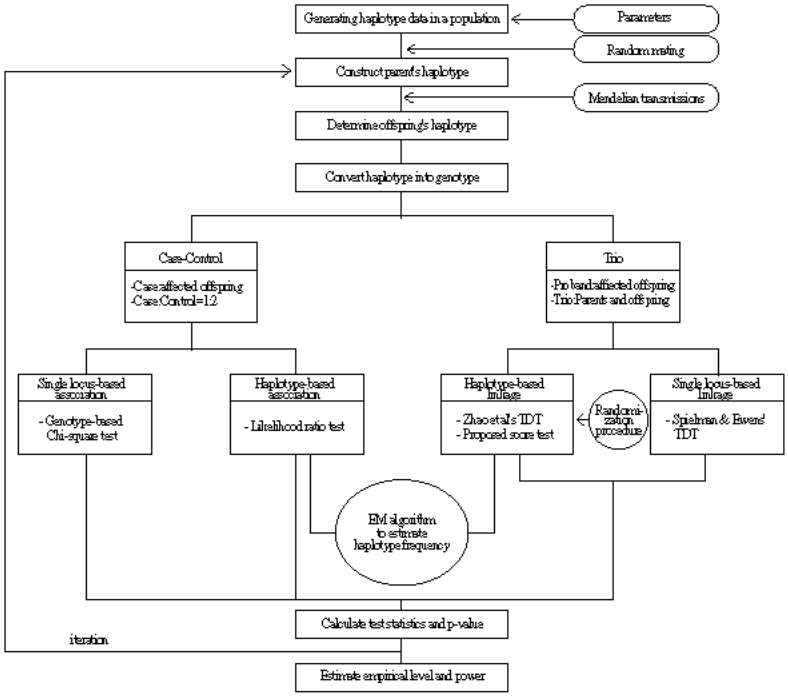

Figure 2. Flow chart of simulation.

에 사용된 RR의 크기는 1.0, 1.2, 1.6, 2.0, 3.0, 4.0, 6.0 이었다.

한편 환자-대조군을 이용한 연관성분석은 인구집단의 혼합 또는 층화에 대해 민감하게 작용하므로 이를 모형에 포함하기 위해서 본 시뮬레이션에서는 연구대상 집단이 2 개의 민족으로 이루어졌다고 가정하였다. 따라서 두 민족의 8가지 일배체형의 비율과 질병발생의 기저위험 비율에 따른 서로 다른 6가지 조합에서 연관성분석과 연쇄분석의 통계적 특성을 살펴보았다 (Table 1). 연구대상자의 유전자 자료발생은 먼저 주어진 모수에 따라 부와 모의 일배체형 짝을 각각 독립적으로 발생한 다음(random mating), 부와 모의 일배체형 짝 중 랜덤하게 하나씩을 선택하고 자녀에게 각각 전달하여 자녀의 유전자형을 결정하였다. 연구대상자 수는 각 민족에서 200명으로 총 400명으로 하였다. 자녀만을 중심으로 환자-대조군 자료를 생성하고 대조군의 수는 환자군의 2배로 하였다. 연쇄분석을 위한 trio 는 질병이 있는 자녀의 부/모/자로 구성하였다.

발생한 유전자형 자료를 토대로 환자-대조군 자료에서는 단일유전자에 대한 유전자형에 기초한 카이제곱 검정과 일배체형을 이용한 LRT(Likelihood Ratio Test)방법, trio 자료에서는 단일유전자에 대한 TDT 통계량, 일배체형에 기초한 Zhao 등 [17]의 TDT 통계량과 본 연구에서 제안한 스코아 통계량을 이용하여 각각 연관성분석과 연쇄분석을 시행하였다. 단, Zhao 등 [17]의 TDT 통계량과 본 연구에서 제안한 스코아 통계량을 계산하는 과정에서 확률화 검정의 횟수를 100번으로 하여 경험적 유의확률을 구하였다. 이러한 과정을 총 200번 반복적으로 시행하면서 유의수준 5%에 대한 각 통계량의 경험적 유의수준과 검정력을 조사하였다. 5% 유의수준에 대한 경험적 유의수준이 (0.0198, 0.0802)의 구간에 있으면 만족스러운 것으로 생각할 수 있다. 이상의 시뮬레이션에 대한 전체 진행도는 Figure. 2와 같다.

결 과

Table 2는 서로 다른 6가지 모형에서 단일유전자의 유전자형을 이용한 카이제곱 검정과 일배체형을 이용한 LRT 방법 (이상 환자-대조군 자료), 단일유전자에 대한 TDT 통계량, 일배체형에 기초한 Zhao 등 [17]의 TDT 통계량 T_{Zhao}과 제안한 통계량 T_s (이상 trio 자료) 에 대한 경험적 유의수준의 결과이다. 환자-대조군의 여러 가지 연관성분석은 Configuration II의 일부를 제외하고는 경험적 유의수준이 명목 유의수준 5%를 대체로 만족하였다. Configuration II에서 세 번째 유전자좌위는 두 민족간에 질병에 대한 기저위험 비율에 차이가 있을 뿐 아니라 m_3 대립유전자의 빈도에도 차이가 있으므로 이의 영향으로 인해 경험적 유의수준이 만족되지 않았다. 또한 세 번째 표지유전자의 영향으로 일배체형을 이용한 LRT 통계량도 경험적 유의수준이 만족되지 않았다. 한편 Configuration II에서 나머지 두 유전자좌위에 있는 대립유전자의 분포는 두 민족간에 동일하여 인구집단의 혼합이 없기 때문에 이 좌위에 대한 연관성분석의 경험적 유의수준이 만족됨을 알 수 있다. 한편 Configuration I에서는 두 민족의 질병에 대한 기저위험 비율에 차이가 없으므로 인구집단의 혼합만으로는 환자-대조군의 연관성분석에 영향을 미치지 않음을 알 수 있다. Configu-

Table 2. Empirical levels of single locus-based and haplotype-based tests according to various configurations

Configuration*	Case-Control				Trio				
	single locus-based chi-square test			haplotype (LRT†)	single locus-based TDT†			haplotype§	
	locus 1	locus 2	locus 3		locus 1	locus 2	locus 3	T_{Zhao}	T_Z
I	0.040	0.045	0.030	0.070	0.030	0.040	0.035	0.055	0.060
II	0.030	0.050	0.270	0.170	0.015	0.035	0.045	0.045	0.050
III	0.065	0.040	0.025	0.045	0.015	0.025	0.015	0.050	0.050
IV	0.055	0.045	0.060	0.060	0.055	0.010	0.045	0.075	0.060
V	0.060	0.050	0.040	0.075	0.035	0.060	0.030	0.035	0.040
VI	0.055	0.070	0.050	0.080	0.045	0.050	0.040	0.035	0.040

* configuration is the same as shown in Table 1.
† likelihood ratio test
‡ transmission/disequilibrium test
§ T_{Zhao} and T_Z are haplotype-based TDT proposed by Zhao et al.(2000) and this study, respectively

Table 3. Empirical powers of single locus-based and haplotype-based tests according to relative risks of disease occurrence for configurations I, IV and V

Configuration*	RR†	Case-Control				Trio				
		single locus-based chi-square test			haplotype (LRT‡)	single locus-based TDT§			haplotype‖	
		locus 1	locus 2	locus 3		locus 1	locus 2	locus 3	T_{Zhao}	T_Z
I	1.2	0.060	0.070	0.055	0.085	0.030	0.050	0.020	0.055	0.065
	1.6	0.080	0.075	0.045	0.140	0.050	0.040	0.010	0.100	0.100
	2.0	0.155	0.125	0.035	0.220	0.090	0.115	0.025	0.115	0.120
	3.0	0.370	0.340	0.065	0.555	0.230	0.195	0.040	0.315	0.300
	4.0	0.665	0.640	0.050	0.820	0.415	0.405	0.040	0.615	0.595
	6.0	0.975	0.975	0.035	0.995	0.760	0.705	0.025	0.960	0.940
IV	1.2	0.070	0.060	0.060	0.055	0.005	0.050	0.035	0.065	0.050
	1.6	0.105	0.090	0.030	0.150	0.070	0.065	0.030	0.110	0.105
	2.0	0.225	0.235	0.030	0.355	0.085	0.115	0.045	0.195	0.195
	3.0	0.565	0.570	0.050	0.840	0.290	0.340	0.050	0.425	0.435
	4.0	0.915	0.910	0.070	0.990	0.500	0.605	0.030	0.800	0.775
	6.0	0.995	1.000	0.060	1.000	0.895	0.805	0.040	0.955	0.960
V	1.2	0.040	0.065	0.030	0.090	0.035	0.060	0.030	0.050	0.060
	1.6	0.115	0.105	0.035	0.185	0.070	0.065	0.025	0.110	0.095
	2.0	0.225	0.245	0.090	0.300	0.110	0.130	0.045	0.170	0.155
	3.0	0.645	0.675	0.055	0.760	0.340	0.365	0.040	0.485	0.495
	4.0	0.940	0.940	0.030	0.990	0.630	0.595	0.045	0.810	0.795
	6.0	1.000	1.000	0.030	1.000	0.920	0.930	0.050	0.985	0.980

* configuration is the same as shown in Table 1
† relative risk is the ratio of the probability of disease occurrence in individuals with disease susceptible alleles to that of disease occurrence in individuals without disease susceptible alleles
‡ likelihood ratio test
§ transmission/disequilibrium test
‖ T_{Zhao} and T_Z are haplotype-based TDT proposed by Zhao et al.(2000) and this study, respectively

ration III 부터 Configuration IV까지는 두 민족간에 모든 유전자좌위의 대립유전자 분포가 동일하므로 연관성분석의 경험적 유의수준이 만족됨을 알 수 있다. 연쇄분석에 관한 TDT 통계량의 경험적 유의수준은 대체로 만족되나 단일 유전자좌위에 대한 방법은 보수적인 경향을 보이고 있다. 그러나 일배체형에 기초한 Zhao 등 [17]의 방법과 본 연구에서 제안한 스코아 통계량은 안정적인 경향을 보였다. TDT 방법은 연구대상자의 인구집단 혼합과 민족간 질병의 기저위험 비율의 차이로 인한 영향을 받지 않음을 알 수 있었다.

Table 3은 Configuration I, Configuration IV, 그리고 Configuration V에서 RR의 변화에 따른 검정력을 비교한 결과이다. 환자-대조군이나 trio 연구설계 모두 단일 유전자좌위를 이용하는 방법보다 일배체형을 이용하는 방법의 검정력이 높았고, 특히 질병관련 유전자의 효과가 커질수록 일배체형을 이용하는 방법의 검정력 증가가 더 높았다. 또한 환자-대조군의 유전자형에 기초한 카이제곱 검정의 검정력이 TDT 통계량의 검정력보다 높았다. 세 번째 유전자좌위에 대한 검정력은 환자-대조군의 연관성분석이나 trio의 TDT 분석에서 유의수준 정도로 매우 낮았다. 이는 질병관련 유전자좌위와 세 번째 유전자좌위 간에 LD가 0으로 평형상태에 있기 때문이다. 이러한 경향은 Configuration I, Configuration IV, 그리고 Configuration V에서 동일하였다. LRT 통계량과 TDT 통계량에서 Configuration I이 Configuration IV 또는 Configuration V에 비해 검정력이 낮은 이유는 상대적으로 환자군 또는 유효한 trio의 수가 Configuration I에서 적었기 때문이다. 환자-대조군의 LRT 통계량은 RR이 2.0 이상인 경우, Configuration V보다 Configuration IV에서 환자군의 수가 상대적으로 많았기 때문에 검정력이 높았다. 그러나 일배체형에 기초한 TDT 통계량은 RR이 커질수록 Configuration IV보다 Configuration V에서 검정력이 높았다. 이러한 이유는 RR이 커질수록 Configuration IV와 Configuration V의 trio 수의 차이는 적어지나, Configuration IV에서 동형접합성의 일배체형이 상대적으로 많아지고, 동형접합성의 일배체형을 갖는 부모의 정보가 TDT 통계량을 계산할 때 제외되므로 유효한 trio의 수가 상대적으로 Configuration V에서 많아졌기 때문이다.

한편 본 연구에서 제안한 스코아 통계량의 검정력은 Zhao 등 [17]의 TDT 통계량의 검정력과 비슷하였다. 그러나 Configuration I에서 RR이 증가할 때 Zhao 등 [17]의 통계량의 검정력이 스코아 통계량의 검정력 보다 약간 높은 경향을 보였다. 전반적으로 RR이 2.0 미만인 경우, 일배체형에 기초한 방법의 검정력은 환자-대조군에서 0.3 미만, trio 연구설계에서 0.17 정도로 낮았다.

고 찰

HapMap 프로젝트는 인간게놈의 DNA 서열상에 존재하는 일배체형 지도를 작성하기 위해 2002년에 시작하였으며, 여러 나라가 참여하는 국제적인 연구로서 여러 민족의 자료를 분석하여 일배체형의 다양성과 일배체형 블록의 크기 등을 밝혔다. 블록 내에 일정빈도 이상인 일배체형(common haplotype)은 평균적으로 3-5개, 블록의 평균 크기는 11-22 kb로서 기대했던 것 보다는 훨씬 인간의 게놈이 단순함을 보였다 [4]. 이러한 사실은 앞으로 질병관련 유전자를 탐색하는 유전체 연구에서

일배체형을 이용하는 것이 현실적으로도 가능하고 앞으로의 연관성분석 방향을 제시하였다는 점에서 큰 의의가 있다. 그러나 HapMap 프로젝트에 대해 부정적인 시각도 많이 있다. 예를 들면 HapMap 프로젝트는 일정빈도 이상인 SNP나 일배체형에 초점을 두기 때문에 만약 common disease - common variant가 아닌 빈도가 낮은 대립유전자에 의해 질병이 발생되는 경우의 일배체형 정보는 유용하지 않을 수 있다 [20]. 이러한 시각의 차이는 유전자를 탐색하는 방법으로서 연쇄분석과 연관성분석 중 어떤 것을 선호하는지에 따른 인식의 차이로 생각할 수 있다. 그렇지만 연쇄분석을 통해 질병관련 유전자의 위치를 개략적으로 찾고, 연관성분석을 통해 위치를 좁힌다는 상호 보완적인 의미로 생각하면 HapMap 프로젝트와 일배체형을 이용한 연관성분석은 큰 의미가 있다고 생각된다.

본 연구는 일배체형을 이용한 새로운 TDT 통계량을 제안하고자 하였다. 새롭게 제안한 방법은, 먼저 Zhao 등 [17]과 동일한 방법으로 자녀에게 전달되는 일배체형의 전달/비전달에 대한 $h \times h$ 분할표를 추정한 후, 추정된 분할표의 행 주변합과 이에 대응하는 열 주변합의 차이에 기초한 스코어 통계량이다. Zhao 등 [17]은 두 주변합 차이의 분산만을 고려한 통계량을 이용하였지만, 본 연구에서는 짝을 이룬 분할표에서 주변합의 차이에 존재하는 공분산도 고려함으로서 더 많은 정보를 통계량에 포함하였다. 따라서 본 연구에서 제안하는 방법이 Zhao 등 [17]의 방법보다 통계량의 점근분포가 더 정확하고 통계적 검정력이 더 높을 것으로 기대되었다. 짝을 이룬 일배체형의 전달/비전달 분할표는 단일 유전자좌위에서 여러 개의 대립유전자를 갖는 경우의 분할표와 형태는 같으나 분할표 내의 각 셀의 자료는 독립이 아니다. 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 각 셀의 독립성을 가정한 카이제곱 점근분포를 사용할 수가 없다. 이러한 문제점을 해결하기 위하여 Zhao 등 [17]의 연구와 본 연구에서는 확률화 검정을 시행하여, 귀무가설 하에서의 경험적 분포를 추정하고 주어진 분할표의 유의확률을 추정하였다.

시뮬레이션을 통한 경험적 유의수준과 검정력을 비교한 결과, 본 연구에서 제안한 방법은 경험적 유의수준이 명목 유의수준 5%를 만족하였으며, Zhao 등 [17]의 TDT 통계량에 비해 검정력의 증가가 없었다. Configuration I에서는 RR이 증가할 때 Zhao 등 [17]의 통계량의 검정력이 본 연구의 스코아 통계량의 검정력 보다 최고 2% 정도 높은 경향을 보였으나 유의수준 5%에 대한 오차의 한계가 약 3%임을 감안할 때 두 통계량의 검정력의 차이는 오차의 범위 내에 속한다고 할 수 있다. 일배체형의 전달/비전달 주변합의 분산-공분산 정보를 Zhao 등 [17]의 통계량에 추가하였으나 검정력의 증가가 없었던 것은 두 방법 모두 통계량의 점근적인 근사분포를 이용한 것이 아니라 확률화 검정을 통해 귀무가설 하에서의 통계량 분포를 유도하였기 때문으로 생각된다. 그러나 두 방법 모두 인구집단의 혼합에 대해 영향을 받지 않았고, 또한 단일 유전자좌위에 기초한 TDT 통계량 보다 검정력이 높았으며, 만약 단일 유전자좌위를 이용한 방법에서 다중검정으로 인한 유의수준의 증가를 조정하면 검정력의 차이는 더욱 커지게 되므로 일배체형을 이용하는 방법이 앞으로 질병관련 유전자를 탐색하는 유전역학 연구에서 주요한 방법으로 사용될 것으로 생각된다.

본 연구에서 사용한 시뮬레이션은 제한된 조건하에서만 이루어졌다. 표지유전자좌위와 질병관련 유전자좌위의 유전적 거리(genetic distance) 또는 물리적 거리(physical distance)를 고려하지 않고, 인구집단의 대립유전자 빈도만을 고려하였으며, 시간이 경과함에 따라 동일한 유전자좌위에서 여러 번의 변이가 발생한 모형을 구축하였다. 이러한 시뮬레이션 모형은 다음과 같은 이유 때문에 고려되었다. 본 연구에서 제안한 TDT 통계량과 Zhao 등 [17]의 통계량은 부/모의 일배체형이 자녀에게 전달될 때 재조합이 없어야 하므로 유전자좌위들의 거리를 자료발생에 대한 모수로 포함하기 어려웠다. 또한 연관성분석 또는 TDT와 같이 연쇄분석과 연관성분석의 의미를 동시에 가지는 방법에서는 질병관련 유전자와 표지유전자 간의 물리적 거리뿐만 아니라 변이가 발생한 시간적인 관계 또는 대립유전자의 이질성(allelic heterogeneity)이 중요할 수 있기 때문이다 [21]. 즉, 세 개의 표지유전자와 질병관련 유전자 사이에 변이가 일어난 시간적인 전후 관계를 주기 위해서는 동일한 유전자좌위에 여러 번의 변이가 발생하는 것이 불가피하였다. 그러나 본 연구에서 사용한 시뮬레이션 모형은 인구집단의 유전적 다형성을 설명하는 coalescent 모형에 근거하고 [22], 질병관련 대립유전자를 가지고 있지 않더라도 환경적 요인에 의해서 질병이 발생할 수 있으며, 그리고 인구집단의 혼합으로 야기될 수 있는 문제점을 포함하였다는 점에서 의의가 있다.

한편 본 연구에서 제안한 스코아 검정과 Zhao 등 [17]의 방법을 이용하여 혈압과 ACE(Angiotension-I Converting Enzyme) 유전자의 관련성을 평가하였으며, Kangwha Study [23] 자료를 사용하였다. Kangwha Study는 1986년 만 6세에 해당하는 초등학생들을 대상으로 혈압, 신체계측과 혈청지질 등에 대해 1997년 17세까지 추적 관찰된 코호트이며, 부모의 혈압 및 이와 관련된 정보도 아울러 조사되어 있다. 본 자료 분석에서는 15세부터 17세까지의 코호트 대상 814명 중 혈압이 같은 연령에 비해 90백분위수 이상으로 상대적으로 높았고, 부모의 혈액시료가 같이 보관되어 있는 40명을 발단자로 하여 부/모/자의 trio 자료를 구축하였다. 조사한 유전자는 17q23에 위치한 ACE 유전자 내의 4가지 SNP(A-240T, C-93T, I/D, A2350G)이며, 발단자에서 4가지 SNP간의 D' 은 모두 1.0으로 높은 LD를 보였다. EM 알고리즘을 이용하여 부/모의 일배체형 빈도를 추정한 결과 6가지 일배체형이 존재하였으며 A-C-I-A와 T-T-D-G

일배체형의 빈도가 각각 53.7%와 41.3%로서 대부분이었으며, 모든 부모의 일배체형이 자녀에게 전달될 때 재조합이 없었다. 총 1000번의 확률화 검정을 시행한 결과, Zhao 등의 검정통계량 값은 7.36 (p=0.143)이었고 본 연구의 스코아통계량은 5.27 (p=0.087)로서 두 검정결과 모두 ACE 유전자와 높은 혈압이 연쇄되어 있다고 보기 어려웠다. 그러나 추정된 일배체형의 전달/비전달 분할표에서 T-T-D-G를 전달하고 다른 일배체형을 전달하지 못한 경우가 29건, 다른 일배체형을 전달하고 T-T-D-G를 전달하지 못한 경우가 17건으로, trio 수의 적음을 고려할 때 한국인에서 T-T-D-G 일배체형과 높은 혈압의 관련성을 계속적으로 연구할 필요성이 있다. 한편 두 검정방법의 유의확률을 비교해 볼 때 본 연구에서 제안한 스코아 통계량의 유의확률이 Zhao 등의 방법보다 작게 추정되어 더욱 다양한 시뮬레이션 환경에서 두 방법의 검정력을 비교할 필요성이 있음을 시사한다.

일배체형을 이용하여 연쇄분석과 연관성분석에 대한 검정력 높은 통계적 방법을 개발하고자 한 본 연구는 부/모의 일배체형 전달에 대한 불확실성을 줄이기 위하여 자녀의 추가 정보를 이용하는 방법, 부와 모 중 한명의 유전형 정보만이 있는 경우, 부/모의 일배체형 전달에서 재조합이 일어난 경우, 표현형이 양적형질(quan-titative traits)인 경우 등 다양한 유전역학 자료에서 응용할 수 있도록 계속 확장해 나갈 예정이다. 또한 시뮬레이션 환경을 확대하고 더욱 다양한 환경에서 각각의 방법들을 비교하여 장단점을 파악할 예정이다.

결 론

본 연구는 일배체형을 이용한 새로운 TDT 통계량을 제안하였으며, 시뮬레이션을 통해 제안한 방법이 단일 유전자좌위를 이용한 TDT 통계량에 비해 검정력이 높음을 보였다. 앞으로 질병관련 유전자를 탐색하는 유전역학 연구에서 일배체형을 이용하는 방법이 활성화 될 것으로 기대되며 또한 본 연구에서 제안한 새로운 방법이 중요하게 사용될 수 있을 것으로 기대된다.

참고 문헌

1. Spielman RS, McGinnis RE, Ewens W. Transmission test for linkage disequilibrium : the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; 52: 506-516
2. Schulze TG, McMahon F. Genetic association mapping at the crossroads: Which test and why? Overview and practical guidelines. *Am J Med Genet* 2002; 114: 1-11
3. Kang D, Lee KM. Current status of genomic epidemiology research. *Korean J Prev Med* 2003; 36: 213-222 (Korean)
4. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. *Science* 2002; 296: 2225-2229
5. The international HapMap Consortium. The international HapMap project. Nature 2003; 426: 789-796
6. Jorde LB. Linkage disequilibrium as a genetic mapping tool. *Am J Hum Genet* 1995; 56:11-14
7. Keavney B. Genetic epidemiological studies of coronary heart disease. *Int J Epidemiol* 2002; 31: 730-736
8. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nat Genet* 2001; 29: 229-232
9. Clark AG. Inference of haplotypes from PCR-amplified samples of diploid population. *Mol Biol Evol* 1990; 7: 111-122
10. Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet* 1995; 56: 799-810
11. Stephens M, Smith NJ, Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; 68: 978-989
12. Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *Am J Hum Genet* 2002; 70: 157-169
13. Zhao JH, Curtis D, Sham PC. Model-free analysis and permutation tests for allelic associations. *Hum Hered* 2000; 50: 133-139
14. Wilson SR. On extending the transmission/disequlibrium test(TDT). *Ann Human Genet* 1997; 61: 151-161
15. Clayton D, Jones H. Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 1999; 65: 1161-1169
16. Clayton D. A generalization of the transmission/disequilibrium test(TDT) for uncertain haplotype transmission. *Am J Hum Genet* 1999; 65: 1170-1177
17. Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenaauer DB, Sun F, Kidd KK. Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* 2000; 67: 936-946
18. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 1996; 59: 983-989
19. Stuart A. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 1955; 42: 412-416
20. Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. *Trends in Genetics* 2003; 19: 135-140
21. Terwilliger JD, Weiss KM. Linkage disequilibrium mapping of complex disease : fantasy or reality? *Curr Opin Biotechnol* 1998; 9: 578-594
22. Rosenberg NA, Nordborg M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev Genet* 2002; 3: 380-390
23. Suh I. Genetic epidemiology study on the long term change of blood pressure and the incidence of hypertension. Korean Research Foundation; 2004 (Korean)