반수체를 이용한
관련성 분석 방법 비교

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
이 은 혜

반수체를 이용한
관련성 분석 방법 비교

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2004년 12월 일

연세대학교 대학원
의학전산통계협동과정
의학통계학전공
이 은 혜

감사의 글

가장 먼저 부족한 저에게 용기를 북돋아 주시고 무사히 논문을 마칠 수 있게 해주신 김동기선생님께 감사드립니다. 유전통계에 대한 끊임없는 호기심을 유발시켜 주셨던 임길섭선생님께도 감사드립니다. 심혈관연구소에서 2년 동안 많은 것을 가르쳐 주시고 유전체 연구에 대한 열정을 몸소 보여주신 박현영선생님께 진심으로 감사를 드립니다.

처음 의학통계 공부를 시작할 때부터 논문을 마치기까지 옆에서 늘 환한 미소로 부족한 나를 이끌고 꾸준히 공부하는 모습으로 나 자신을 채직질하게 해준 찬미언니에게 마음속 깊은 고마움을 전합니다. 지난 2년간 심혈관유전체연구소에서 매일매일 함께 있어 즐거움과 든든함을 안겨준 지누언니, 은정언니, 홍정애선생님, 이뿐선영 그리고 수진언니, 정신영선생님께 감사한 마음과 함께, 같이있어 행복했다는 말을 전하고 싶습니다. 늘 연구하는 모습으로 연구하는 사람의 본이 되어주신 조은영선생님께 감사드립니다. 이제 남은 겨울을 함께할 상미언니, 최정란선생님의 따뜻한 위로도 많은 힘이 되었습니다. 감사합니다.

의학통계에 대한 끊임없는 고민을 하게 한 기준오빠에게 감사드립니다. 많은 것을 같이 하지는 못했지만 성민오빠, 미영언니, 무영오빠, 원열오빠 에게도 감사의 마음을 전합니다. 혜리언니의 멋진 논문도 기대하겠습니다. 많이 챙겨주지는 못했지만 귀여운 소연, 민진, 성은도 앞으로의 대학원 생활을 잘하리라 믿습니다.

지난 2년 동안 나를 붙들어 주고 위로하고 힘이 되어준 사랑하는 동기 수옥언니와 신영언니.. 대학원 과정에서 동기로 만나 같이 공부한 것이 내게 행운이었습니다. 앞으로 좋은 일만 함께 했으면 좋겠습니다.

가장 아름다운 시절을 함께했던 나의 사랑하는 대학동기 부르주아-현미, 정아, 지원, 은영, 그리고 귀여운 아름, 영심에게 사랑과 고마움을 전합니다. 힘든 시기에 같이 행복했던 시간을 떠올리는 것만으로도 많은 힘이 되었습니다. 언제든지 웃음으로 나에게 힘을 넣어준 친구 동구, 인희에게도 고맙다는 말을 하고 싶습니다. 항상 내 편에 서준 인천 대일 친구들의 격려도 지칠 때마다 버틸 수 있는 힘을 안겨주었습니다.

누구보다도 가장 사랑하고 감사드리는 분은 나를 믿고 지원해 주시는 부모님입니다. 많은 걱정을 끼쳐드리는 큰 딸임에도 불구하고 항상 믿음과 사랑으로 감

싸주시고 그 안에서 스스로 일어나게 하는 부모님께 표현 못할 만큼의 감사와 사랑을 드립니다. 또한 나의 아름다운 동생 지혜에게 지혜가 언니에게 주는 사랑보다 더 많이 사랑한다는 말을 전하고 싶습니다.

마지막으로 힘든 시기의 시작부터 논문을 마감하기까지 옆에서 쉼 없는 조언과 위로, 사랑을 주며 같이 기뻐하고 슬퍼해준 소중한 그대에게 진심 어린 사랑과 고마운 마음을 전합니다.

2004년 12월

이 은 혜 올림

<목 차>

<표 차례>

<그림 차례>

국 문 요 약

반수체를 이용한 관련성 분석 방법 비교

인간유전체 연구에서는 단일 표식유전자(single marker)에서의 유전형(genotype)과 형질간의 관련성분석을 통하여 질병과 같은 복합형질(complex trait)과 관련된 유전자를 발견하는 것에 관심을 가져왔다. 그러나 군집되어있는 유전자들이 질병에 영향을 줄 수 있기 때문에 단일 표식유전자의 정보만 이용하여 관련성을 찾는 것은 쉽지 않고 검정력이 낮은 경향이 있다. 따라서 여러 SNP들의 정보를 포함하고 있는 반수체(haplotype)를 이용한 분석은 보다 효율적으로 관련성을 검정할 수 있다. 반수체와 형질간의 관련성을 검정하는 방법은 실험군과 대조군간의 반수체 빈도(haplotype frequencies)를 비교하는 것이 일반적인 것이었다. 최근에는 회귀모형에 기초한 방법(regression based method)을 사용하여 이분형형질(binary trait) 뿐 아니라 양적형질(quantitative trait)에 대한 반수체 관련성 분석을 할 수 있으며, 비유전적 변수(nongenetic covariates)의 영향을 고려한(adjusted) 분석도 가능하게 되었다.

본 논문에서는 선형모형에 기초한 반수체와 양적형질간의 관련성분석 방법으로 스코어 통계량을 이용하는 방법(score test)과 HTR 방법(haplotype trend regression)을 심혈관유전체자료에 적용하고 결과를 비교하였다. 다양한 크기의 반수체와 양적형질들과의 관련성분석 결과, 유의성이 존재할 때 HTR 방법이 스코어 방법보다 높은 검정력(power)을 보였다. 반수체를 구성하는 SNP수가 늘어남에 따라 HTR 방법 결과와 스코어 방법 결과의 차이가 증가하는 경향을 보였다.

핵심되는 말 : 반수체, 관련성분석, 스코어 방법, HTR 방법, 양적형질

제 1 장 서 론

1.1 연구 배경

최초 인간 유전체에 대한 정보의 확대와 더불어 유전성질환 뿐 아니라 당뇨나 심장병, 고혈압 등과 같은 복합 형질에 영향을 미치는 유전자 발견에 많은 관심이 모아지고 있다. 복합형질은 몸무게, 나이 등과 같은 비유전요인(non-genetic factor)과 유전요인(genetic factor), 환경요인(environmental factor)등 다중요인(multi-factorial)의 영향을 받는다. 질병을 일으키는 유전요인을 가지고 있다고 해서 모두 질병이 발생하는 것이 아니며, 같은 질병 현상이 나타나도 다른 비유전요인에 의한 것일 수 있다. 유전요인과 환경요인이 동시에 질병에 영향을 미치기 때문에, 관찰된 표현형(phenotype)과 질병 발생을 일으키는 유전형(genotype)의 관계는 명확하게 정의되지 않으며 멘델의 유전법칙을 따르지 않는 경우가 많다.

혈연관계가 없는(unrelated) 모집단에서 형질에 영향을 미치는 유전자 발견을 위하여 기본적으로 단일 표식유전자와 복합형질간의 관련성(association)을 검정하는 방법이 제시되었다. 그러나 여러 개의 단일 표식유전자가 밀집되어 형질에 영향을 미치는 것일 수 있으므로 단일 표식유전자를 이용한 형질과의 관련성 검정 방법은 정보력이 낮고 관련성 있는 유전자를 찾기 힘들다. 반수체는 인접한 표식유전자들의 묶음으로, 반수체를 이용한 관련성분석은 다중 표식유전자 정보(multiple marker information)를 포함하고 있기 때문에 단일표식유전자를 사용한 관련성분석보다 검정력이 높고 로버스트한 분석을 할 수 있다 (Akey et al, 2001).

반수체를 이용한 관련성 검정의 고전적인 방법은 환자군과 정상군에서 추정된 반수체의 기대빈도(expected frequency)로 구성된 분할표(cross-table)에 대하여 카이제곱검정을 하는 것이다(Fallin et al, 2001). 분할표를 이용한 집단 간 비교 방법에서는 이분형 형질에 대한 관련성 분석만 가능하다는 제한점이 있으며 유전요인 이외에 환경요인에 대한 고려를 할 수가 없다. 이러한 점을 극복하는 방법으로

선형모형을 이용한 형질과 반수체의 관련성 분석 방법이 제안 되었다. 선형 회귀모형을 사용하면 질적 형질 뿐 아니라 지질 농도나 혈압 등 양적 형질과 반수체의 관련성 분석도 가능하다. 통계적으로 추정된 반수체 빈도를 이용하여 베이지안 방법을 통해 가능한 반수체 조합에 대한 사후확률(posterior probability)을 구할 수 있으며, 이 사후확률은 회귀모형에서 가중치(weight)로 사용되어 가중치준 회귀모형을 이용한 반수체 관련성 분석을 하게 된다. 선형모형을 이용한 반수체 관련성분석 방법에는 스코어 통계량을 이용하는 방법과 HTR 방법이 있다.

1.2 연구 목적 및 내용

본 논문에서는 반수체와 양적형질 간의 관련성을 분석하는 방법으로 선형모형을 이용한 방법들에 대해 비교를 한다. 선형모형을 이용한 반수체 관련성 분석 방법으로 스코어 방법과 HTR 방법을 사용하며, 심혈관계유전체연구센터에서 수집된 유전체자료를 이용하여 두 방법을 통한 실제 양적형질과 반수체의 유전적 관련성분석을 수행하고 비교한다.

반수체 정보를 이용한 관련성 분석은 두 단계 과정으로 이루어진다. 먼저 반수체 빈도를 추정하고, 그 다음 추정된 반수체 빈도와 형질의 관련성이 존재하는지 검정 한다.

혈연관계가 없는 개인(unrelated individual)에서 유전형에 대한 정보만 가지고는 어떠한 조합으로 반수체가 구성되어 있는지는 알 수 없으므로(ambiguous) 통계적인 방법을 사용하여 반수체 빈도를 추정한다. 반수체 빈도를 추정하는 통계적 방법으로는 클락 알고리즘(Clark, 1990)과 EM 알고리즘(Excoffier and Slatkin, 1995), 베이지안 방법(Stephen et al, 2001)이 있다. 그 다음 추정된 반수체 빈도를 양적형질과 연관시켜 반수체와 양적형질간의 관련성이 존재하는지 분석하게 된다.

스코어 방법과 HTR 방법에서는 EM 알고리즘을 통하여 반수체 빈도를 추정하

며, 스코어 방법에서는 일반화 선형 모형을 통하여 얻은 스코어 통계량을 이용하고 HTR 방법은 단순선형모형을 통하여 형질과 반수체의 관련성을 검정한다. 이 연구결과는 선형모형을 기반으로 하는 관련성 분석 방법인 스코어 방법과 HTR 방법의 차이를 보여주며, 양적형질의 반수체 관련성 분석을 계획하고 평가할 때 유용하다.

논문의 제 1장 서론에서는 연구의 배경이 되는 양적형질과 양적형질의 유전자 관련성분석에 대해 소개하고 연구 내용을 제시한다. 제 2장에서 반수체 관련성 분석을 위한 첫 단계인 반수체 빈도를 추정하는 방법에 대해 간단히 설명하고 반수체 빈도를 비교하는 고전적 방법에 대해서 기술한다. 또한 선형모형을 기반으로 하는 반수체 관련성 분석 방법인 스코어 방법과 HTR 방법에 대해서 기술한다. 스코어 방법과 HTR 방법에 대한 이론적 비교내용은 제 3장에서 기술한다. 제 4장에서는 두 방법의 통계적인 성능을 비교하기 위해 심혈관 자료에의 적용 과정에 대해 설명하고 연구 결과를 제시한다. 제 5장에서는 연구 결과에 대한 결론을 발표한다.

제 2 장 반수체를 이용한 관련성 분석

2.1 반수체 빈도 추정

2.1.1 반수체의 모호성

반수체는 하나 이상의 위치(locus)에서 부모로부터 유전된 대립유전자(alleles)로 이루어진 특정한 조합이다. 인간은 어머니와 아버지로부터 두 개의 반수체를 유전 받으며, 어떠한 조합으로 유전됐는지 알 수 있다면 다중위치유전형(multi-locus genotype)이나 반수체를 분명하게 정할 수 있다. 그러나 대부분의 경우에 각 위치(locus), 즉 단일위치유전형(single-locus genotype)으로 존재하는 대립유전자는 알고 있지만 반수체가 어떤 조합으로 구성되었는지 분명하게 정하지 못하며 이것을 반수체가 모호하다(haplotype ambiguous)고 표현한다.

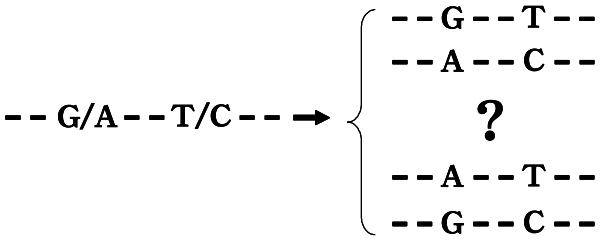

그림 1. 반수체 모호성(haplotype ambiguous)
첫 번째 위치의 대립유전자는 G와 A이고 두 번째 위치의 대립유전자는 T와 C일 때, 두 가지의 반수체를 구성할 수 있으며 유일한 반수체를 결정하지 못한다.

예를 들어서, 한 사람이 어떤 위치에 대립유전자 A 와 a를 가지고 있고, 다음 위치에는 대립유전자 B 와 b를 가진다고 하자. 이 때 관찰된 단일 유전형은 Aa, Bb 이고, 유전되는 반수체는 AB와 ab (AB/ab)인 경우와 Ab와 aB (Ab/aB)인 두

가지 경우가 있다. 구성 가능한 반수체의 수는 각 위치에 있는 대립유전자 수의 곱이 된다. 두 개의 대립유전자(biallelic)를 갖는 표식자가 m 개 존재 한다면 가능한 반수체는 2^{m} 개 이다.

단일위치유전형이 이질적(heterozyguous)인 위치가 많아질수록 구성될 수 있는 다중위치유전형의 수가 많아지는데, 이질적인 단일위치유전형 위치가 $h > 0$ 개 이라면 구성되는 다중위치유전형의 수는 2^{h-1} 개가 된다. 예를 들어 $h = 1$ 일 때, 세 개 위치 A, B, C에서 C 위치(locus)에서만 유전형이 이질적이라면 다중위치유전형은 $a_1b_1c_1/a_1b_1c_2$ 으로 나타낼 수 있다. $h = 2$ 일 때, 위치 B 와 C의 유전형이 이질적이라면 $a_1b_1c_1/a_1b_2c_2$, $a_1b_1c_2/a_1b_2c_1$인 2 가지 경우의 다중위치유전형이 존재하게 된다.

반수체는 고급기술을 사용하거나 가족 구성원의 유전 자료 수집을 통해서 알아낼 수 있다 (Judson and Stephens, 2001). 부모의 유전형을 알면 유전되는 대립유전자를 결정 할 수 있으므로 다중위치유전형을 추정할 수 있는 것이다. 그러나 이러한 방법들은 부모가 존재하지 않거나 부모의 유전형을 알 수 없는 경우에 사용할 수 없고 비용 면에서 현실적으로 조사 불가능한 경우가 많다. 이에 대한 대안으로 효과적으로 통계적으로 추정해내는 여러 가지 방법이 제안되었고 현재 반수체 연구에 널리 사용되고 있다.

혈연관계가 없는 개인에서 반수체 빈도를 추정하는 통계적 방법으로 몇 가지 알고리즘이 존재한다. 첫 번째 방법은 클락 알고리즘으로, 유전형자료에서 반수체의 종류가 최소가 되도록 이미 알고 있는 반수체를 시작으로 하여 그 다음 올 수 있는 반수체들을 추정 하는 방법이다(Clark, 1990). 다음으로 EM(expectation maximization) 알고리즘을 사용하여 반수체 빈도를 추정하는 방법(Excoffier and Slatkin, 1995)이 있으며 자세한 내용은 다음 절에서 소개한다. 베이지안 방법(Bayesian method)은 유전 이론을 이용하여 반수체의 사전 분포(prior distribution)를 모형화 함으로서 반수체 빈도를 추정하는 방법이다(Stephens et al, 2001).

2.1.2 EM 알고리즘을 이용한 반수체 빈도 추정

클락 알고리즘을 이용한 반수체 빈도의 추정은 반수체를 구성하는 계산 과정이 효율적이며 처리 시간은 빠르지만, 알고리즘이 시작되기 위해서는 명확히 정해진 반수체를 가지고 있는 개인이 존재해야 하고 놓여진 유전형의 순서에 따라서 추정하는 반수체가 달라진다는 문제점이 있다. Excoffier은 EM 알고리즘을 통한 반수체 추정 방법을 제안했다(Excoffier and Slatkin, 1995). EM 알고리즘은 로그 우도(log likelihood) 함수를 최대로 만들기 위하여 상대도수 추정치를 반복적으로 추정하는 방법이다. EM 알고리즘을 사용하면 유전형 자료에 결측값이 존재할 때 이를 추정해낼 수 있으며 각 반수체 쌍에서 사후 확률을 구할 수 있다.

반수체 빈도를 추정하는 EM 알고리즘은 다음과 같이 진행된다. 먼저 n 명의 표본에서 로그 우도 함수는

$$\ln L = \sum_{i=1}^{n} \ln p_i$$

로 나타낼 수 있다. p_i 는 i 번째 사람에서 관찰된 유전형의 확률(genotype probability)로

$$p_{i=} \sum \Pr [\, h_k / h_l \,]$$

와 같이 단일 유전형들로부터 만들어질 수 있는 모든 다중위치유전형의 확률을 합한 값이다. h_k / h_l 은 반수체 k 과 l로 이루어진 다중위치유전형이다. γ_j 가 반수체 j에 대한 확률이라고 할 때, 다중위치유전형의 확률은

$$\Pr[\ h_k / h_l\ ] = \begin{cases} \gamma_l^2 & \text{if} \quad k = l \\ 2\gamma_l\gamma_k & otherwise \end{cases}$$

로 나타낼 수 있다.

i 번째 사람에서 h_l 반수체들의 수를 N_{l_i} 라고 할 때, t 번 째 반복 과정에서의 EM 알고리즘은 다음과 같다.

1. E step : 각 반수체의 기대수(expected number)를 계산한다.

   i번 째 사람에서 관찰된 유전형에 대한 다중위치유전형을 h_k / h_l 라 할 때, 반수체 h_l 의 기대수는

   $$n_{l_i} = E[\ N_{l_i} \mid genotype\ \ of\ \ individual\ \ i\ ] = \frac{2\gamma_l^{(t)}\gamma_k^{(t)}}{p_i^{(t)}} \quad \text{이고,}$$

   $$n_l = E[\ N_l \mid data] = \sum_{i=1}^{n} n_{l_i}$$

   와 같이 구할 수 있다.

2. M step : 계산된 반수체의 기대수를 이용하여 반수체 빈도의 최대 우도 추정치를 찾는다.

   $$\gamma_l^{(t+1)} = \frac{n_l}{2n}$$

위치의 수나 대립유전자의 수가 증가하면, 만들어질 수 있는 반수체 수도 크게 증

가한다.

2.1.3 고전적 반수체 관련성 분석

혈연관계가 없는 표본에서 질병에 걸린 집단과 걸리지 않은 집단 간에 반수체 빈도의 차이가 존재하는지를 검정해서 두 집단간에 유의한 차이가 있으면 그 반수체가 질병을 일으키는 유전자와 관련되어 있다고 설명할 수 있다. 분할표 카이 검정은 두 집단간의 반수체 빈도의 차이가 있는지를 검정하는 방법 중 하나이다. 이러한 고전적 관련성분석 방법은 실험-대조군 분석이나 코호트 조사로부터 나온 자료에서 질병 발생에 영향을 미친다고 의심되는 부분(candidate region)의 유전형과 질병 간의 관련성을 보는 대표적인 검정방법이다.

질병에 걸린 군과 대조군에서 반수체의 빈도를 비교하는 방법으로 우도비통계량(likelihood-ratio statistic)을 사용하는 방법이 있다. 질병에 걸린 집단에 대한 로그 우도 함수를 $\ln(L_{case})$, 대조군에서의 함수를 $\ln(L_{control})$, 전체에서의 함수를 $\ln(L_{total})$ 라고 하고, EM 알고리즘을 사용하여 로그 우도 함수를 최대화 시킨다. 우도비통계량은 다음과 같으며,

$$LR = 2[\ln(L_{case}) + \ln(L_{control}) - \ln(L_{total})]$$

근사적으로(asymptotically) χ^2 분포를 따른다.

위와 같은 χ^2 검정 방법은 유전요인과 연관시킬 형질이 질병의 유무를 나타내는 이분형 형질로 제한되며 각 반수체(specific-haplotype)에서 형질과의 관련성 결과도 보여주지 못한다. 이러한 단점을 보완하기 위하여 선형모형(regression model)을 이용한 방법이 제안되었다.

2.2 스코어 방법

스코어 방법은 일반화 선형 모형을 사용하여 다양한 형태의 형질에 대해서 반수체의 관련성 분석이 가능하도록 하였으며, 비유전요인이 형질에 미치는 영향을 고려한 통계량을 구할 수 있다. 스코어 검정의 또 다른 특징은 형질과 반수체 전체에 대한 관련성 검정(global test)과 각 반수체에 대한 검정이 가능하다는 것이다.

2.2.1 일반화 선형 모형

일반화 선형 모형(generalized linear model)은 선형 방정식을 통하여 반응 평균의 어떠한 함수와 설명변수들의 관계를 규명하고 설명변수의 영향을 평가하기 위한 모형이다. 일반화 선형 모형을 사용하면 연속형, 이분형, 순서형과 같은 다양한 형태의 형질에 대한 선형모형을 세울 수 있으며, 형질의 분포에 따라서 적합한 모형을 생각할 수 있다(Schaid, 2002). 스코어 방법에서는 환경요인의 영향과 유전요인의 영향을 동시에 고려한 관련성 검정이 가능하다.

양적형질의 평균을 η 라 하고 X_e 와 X_g 를 각각 환경적 요인과 유전적 요인에 대한 벡터라 할 때,

$$\eta = X_e'\alpha + X_g'\beta$$

와 같은 선형 모형으로 표현할 수 있다. α와 β는 환경요인과 유전요인에 대한 회귀계수로 각 요인의 효과 정도를 나타낸다. 유전요인과 형질의 관련성 분석에서 귀무가설은 유전요인과 형질간의 관련성이 존재하지 않는다 이며, 유전요인의 효과가 없다는 것으로

$$H_o: \beta = 0$$

과 같이 표현한다. 벡터 Z 와 γ 를 각각 $Z=(X_e \mid X_g)$ 와 $\gamma=(\alpha \mid \beta)$ 라고 할 때, 벡터 Z 에 대한 형질 y 의 우도비 함수를 일반화 선형 모형으로 아래와 같이

$$L(y \mid Z)=\exp\left[\frac{y\eta - b(\eta)}{a(\phi)} + c(y,\phi)\right]$$

나타낼 수 있다(McCullagh and Nelder 1983). 여기서 φ 는 분산(dispersion)을 나타내는 모수이며, 함수 a, b, c는 아래 표에 나타난 대로 형질의 분포에 따라서 알 수 있다.

표 1. 여러 분포에 따른 일반화 선형 모형의 함수들

분포	$\tilde{y}$	$a(\phi)$	$b''(\eta)/a(\phi)$
정규분포	η	σ^2_{mse}	$1/\sigma^2_{mse}$
이항분포	$e^{\eta}/(1+e^{\eta})$	1	$\tilde{y}(1-\tilde{y})$
포아송분포	e^{η}	1	$\tilde{y}$

2.2.2 모호한 반수체에 대한 스코어 검정

환경요인과 유전요인에 대한 벡터 $Z=(X_e \mid X_g)$ 의 스코어 통계량은

$$U_\gamma = \sum_{i=1}^{N} \frac{\partial \ln(L_i)}{\partial \gamma} = \sum_{i=1}^{N} \frac{y_i - \tilde{y}_i}{a(\phi)} Z_i$$

으로 나타낼 수 있다. $\tilde{y}_i$ 는 벡터 Z_i 에 의해 추정된 값이며 γ 는 회귀 모수이다. $a(\phi)$ 는 가정한 형질의 분포에 따라 위의 표에서 선택한다. 스코어 방법에서는 환경요인을 제어한 유전요인의 관련성을 분석할 수 있다. 먼저 유전요인에 대한 영향 β 를 0으로 하고 환경요인에 대해서만 회귀분석을 하여서 환경요인의 영향 $\hat{\alpha}$ 를 추정한다. 추정된 환경요인의 영향 $\hat{\alpha}$ 에 대해서 $U_{\hat{\alpha}} = 0$ 이 되도록 $\tilde{y}_i$ 를 결정하면 형질에 대해서 유전요인에 대한 영향만을 생각할 수 있게 된다. 반수체가 모호할 때 환경적 영향을 제어한 반수체와 형질의 관련성을 보는 스코어 통계량은

$$U_\beta = \sum_{i=1}^{N} \frac{y_i - \hat{y}_i}{a(\phi)} E_p(X_{gi})$$

와 같다. $E_p(X)$ 는 표식유전자 자료가 주어졌을 때 반수체와 형질의 관련성이 없다는 귀무가설 하에서 계산한 유전형 사후분포의 기대값이다. 기대값은 다음과 같이 나타낼 수 있으며,

$$E_p(X) = \sum_{g \in G} X_g Q(g)$$

이 때 유전형의 사후확률 $Q(g)$ 는 $Q(g) = P(g) / \sum_{g \in G} P(g)$ 와 같이 구한다. $P(g)$는 앞서 EM 알고리즘으로 추정한 반수체 빈도의 확률이다. 통계량 U_β 에 대한 분산은

$$V_\beta = V_{\beta\beta} - V_{\beta\alpha} V_{\alpha\alpha}^{-1} V_{\alpha\beta}$$

로 나타낼 수 있다. $V_{\alpha\alpha}, V_{\alpha\beta}, V_{\beta\beta}$는 공변량에 대한 행렬로 다음과 같이 구한다.

$$V_{\alpha\alpha} = \sum_{i=1}^{N} \frac{b''(\eta_i)}{a(\phi)} X_{ei} X'_{ei}$$

$$V_{\alpha\beta} = \sum_{i=1}^{N} \frac{b''(\eta_i)}{a(\phi)} X_{ei} E_p(X'_{gi})$$

$$V_{\beta\beta} = \sum_{i=1}^{N} [\frac{b''(\eta_i)}{a(\phi)} - \frac{(y_i - \tilde{y}_i)^2}{a(\phi)^2}] \; E_p(X_{gi} X'_{gi})$$

$$+ \frac{(y_i - \tilde{y}_i)^2}{a(\phi)^2} E_p(X_{gi}) E_p(X'_{gi})$$

스코어 방법의 또 다른 장점은 전체적 스코어 통계량과 각 반수체에 대한 스코어 통계량을 구할 수 있다는 것이다. 전체적 스코어 통계량은 다음과 같으며,

$$S = U_\beta' \; V_\beta^{-1} U_\beta$$

표본 크기가 클 때 자유도가 V_β의 계수(rank)인 χ^2 분포를 따르게 된다. 스코어 방법은 각 반수체에 대해서 형질과의 관련성을 볼 수 있는데 단일 반수체에 대한 효과가 전체 반수체를 사용한 효과보다 크다면 전체적인 스코어 통계량을 사용하는 것보다 단일 반수체와 형질의 관련성을 보는 분석 결과가 검정력이 커지게 것이다. k 번째 반수체에 대한 스코어통계량은

$$z_k = U_{\beta,k} / \sqrt{V_{\beta,k,k}}$$

이며 자유도 1인 χ^2 분포를 따른다.

스코어 통계량을 이용한 방법은 최대 우도 추정치를 계산할 필요가 없으므로 우도비 검정통계량에 비해서 계산하는 시간이 빠르다.

2.2.3 반수체 스코어링

반수체를 스코어링 하는 방법은 몇 가지가 있다. 먼저 가장 많이 쓰이고 간단한 방법은 가지고 있는 반수체의 수를 세는 것으로 0, 1, 2 값을 부여하는 방법이다. 다른 방법은 형질과 관련성이 있다고 여겨지는 염색체 부분을 훑어 나가는 방법으로 “window sliding" 스코어링 방법 이라고 한다. Window sliding은 인접한 위치에 있는 유전자들을 순서대로 일정한 단위(window)로 묶어서 질병과 관련되어 있다고 의심이 되는 부분을 훑어가는(sliding) 방법이다. 묶는 유전자 수를 늘려가면서 sliding을 시킬 수 있으며 반수체의 어떤 부분과 형질과의 관련성을 분석한다.

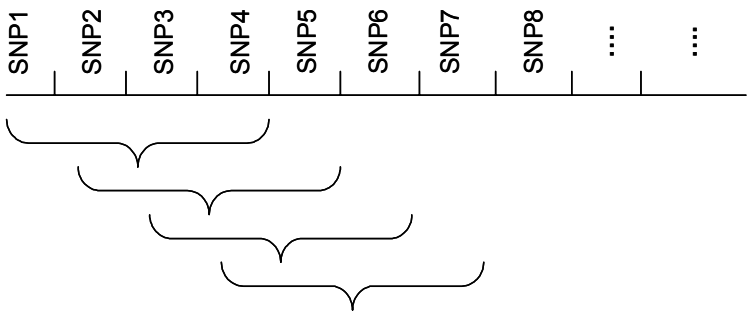

그림 2. window sliding 방법의 예
window 크기를 4로 했을 때, 앞에서부터 차례로 window를 sliding시키면서 관련성분석을 한다.

또 다른 방법으로는 반수체를 구성하는 대립유전자가 구성된 방향에 따라 차이가 있는지에 관심이 있을 때 1과 -1값을 주는 스코어링 방법이다. 예를 들어, 첫 번째 위치의 유전형이 A와 a이고 두 번째 위치의 유전형이 B와 b일 때, 다중위치유

전형이 AB/ab 일 때가 Ab/aB 일 때보다 형질과의 관련성이 더 강하게 존재하는지 보고자하면 관심이 있는 경우에 1을, 반대의 경우에 -1의 값을 주어 차이를 본다.

2.3 HTR(haplotype trend regression) 방법

혈연관계가 없는 개인에서 선형모형을 이용하여 반수체와 형질간의 관련성을 분석하는 다른 방법으로 HTR 방법이 있다. EM 알고리즘을 이용하여 반수체 빈도를 추정하고 추정된 반수체 사후확률과 양적형질의 관계를 단순선형모형을 사용하여 접근한다.

2.3.1 HTR 방법 개요

반수체 관련성 분석에서 실험 대조군 모형을 $2 \times L$ 테이블로 나타낼 수 있다. 테이블에서 두 행은 질병의 유무를 나타내며 L 개의 열은 관찰된 유전형으로 만들어진 반수체를 나타낸다. 이 분석은 두 집단간의 대립유전자의 빈도가 차이 있는지를 보기 위한 분할표 카이검정과 비슷하며, 우도비 검정이나 피어슨의 적합도 검정(Pearson's goodness-of-fit test)과 유사한 방법을 사용한다. χ^2 적합도 검정통계량은 단순선형함수에서 종속변수가 이분형일 때 분산분석(ANOVA)으로 얻어진 F 통계량과 같다. 테이블에 들어가는 값은 EM 알고리즘을 통해서 추정된 반수체의 수이며, 표본크기가 N 일 때 테이블의 값을 모두 합하면 $2N$ 이 된다. 실험군과 대조군으로 들어가는 개인에서 반수체가 두 개 측정되기 때문이며 이렇게 표본이 두 번 세어지게 된다. 종속변수가 질병의 유무를 나타내는 이분형이 아니라 연속형 자료 일 때 같은 방법을 적용한다면 2 배 표본크기에서의 분산분석 방법을 사용 한 것이 된다.

HTR 방법은 2 배 표본크기에서의 분산분석 방법의 대안으로, 표본크기를 고정시킨 분산분석 방법이다. 각 개인이 유전 받는 반수체들의 사후 기대수(posterior expected number)를 독립변수로 사용하여 단순회귀모형을 적용한다. 이 방법은 2배 표본에서의 분산분석보다 더 높은 검정력을 갖는다(Zaykin et al., 2002).

2.3.2 HTR 모형

X 를 L 개 반수체 중 하나라고 정의할 때, 표본크기 N 에서 i 번째 사람이 가지고 있는 반수체는 X_{i1} , X_{i2} 로 나타낼 수 있으며, 이 때 형질을 Y_i 로 나타낸다. 반수체와 형질의 관련성을 검정하기 위해 분산분석 모형은 $Y_{(d)} = A\alpha + \varepsilon$ 로 세울 수 있다. A_{ij}^{T} 는 i 번째 사람이 반수체 j 를 갖는지 아닌지를 나타내는 1 × L 벡터이며 반수체에 대한 정보를 나타내는 벡터를 $A^{T} = (A_{11}, A_{12}, A_{21}, A_{22}, \ldots, A_{N1}, A_{N2})$ 로 표현한다. 예를 들어 $X_{ij} = 2$ 라고 한다면, 반수체 j 가 두 번째 종류의 반수체라는 것을 의미하고 $A_{ij}^{T} = (0\ 1\ 0\ \ldots\ 0)$ 로 나타낸다. 이 때의 형질에 대한 벡터는 $T_{(d)}^{T} = (Y_1, Y_1, Y_2, Y_2, \ldots, Y_N, Y_N)$ 이며 표본의 수가 두 배된 분석을 하게 된다.

HTR 방법은 표본수를 고정시킨 N 차원의 회귀분석을 하는 방법으로, 두 배 표본을 쓰는 분산분석 방법보다 검정력이 높아진다. N 차원 회귀 모형은 다음과 같은데,

$$Y = D\beta + \varepsilon$$

형질은 $Y^T=(Y_1, Y_2, \ldots, Y_N)$ 로 나타내며 반수체에 대한 벡터는 $D^T=(D_1, D_2, \ldots, D_N)$, i 번째 사람의 반수체에 대한 벡터는 $D_i^T=(D_{i1}, D_{i2}, \ldots, D_{iL})$ 로 나타낼 수 있다. 여기서 D_{ij} 는 다음과 같이 정의할 수 있다.

$$D_{ij}=\begin{cases}1 & \text{if } i\text{th subject is homozygous for haplotype } j\\ \frac{1}{2} & \text{if } i\text{th subject is heterozygous including haplotype } j\\ 0 & \text{otherwise}\end{cases}$$

예를 들어, 형질 Y_1 , Y_2 , Y_3 을 갖고 모호하지 않은(unambiguous) 반수체 쌍이 h_1/h_1 , h_2/h_3 , h_1/h_3 일 때, HTR 방법을 사용한 회귀 모형 $E(Y)=D\beta$ 를 생각해 보자.

$$E\begin{pmatrix}Y_1\\ Y_2\\ Y_3\end{pmatrix}=\begin{pmatrix}1 & 1 & 0 & 0\\ 1 & 0 & 1/2 & 1/2\\ 1 & 1/2 & 0 & 1/2\end{pmatrix}\begin{pmatrix}\mu\\ \beta_1\\ \beta_2\\ \beta_3\end{pmatrix}$$

$$\begin{matrix}\quad & \mu & h_1 & h_2 & h_3\end{matrix}$$

반수체가 모호하다면 D 행렬의 2번째에서 4번째 열의 값은 더 이상 0, 1, $\frac{1}{2}$ 값을 가지지 않으며, 유전형이 주어졌을 때의 조건부 확률로 추정해야 한다. 예를 들어 유전형 G_i 를 갖는 i 번째 사람이 갖는 반수체 쌍 (h_2, h_3) 을 갖

는다고 할 때, 반수체 h_2 와 h_3 을 갖게 될 유전형에 대한 조건부 확률은

$$\Pr(h_2, h_3 \mid G_i) = \frac{\Pr(G_i \mid h_2, h_3)\, p_{h_2} p_{h_3}}{\sum_{u,v} \Pr(G_i \mid h_u, h_v) p_{h_u} p_{h_v}}$$

로 나타내며, 여기서 p_{h_u} 와 p_{h_v} 는 반수체 빈도를 말한다. 양적형질과 반수체의 관련성 분석은 반수체와 형질간의 관련성이 없다는 귀무가설 $H_0 : \beta = 0$ 에 대해서 F 검정을 하게 된다. F 검정 통계량은

$$F = \{SSA/(L-1)\}\{SSE/(N-L)\}$$

로 구할 수 있으며 SSA 와 SSE 값은 다음과 같이 구한다.

$$SSA = Y^T (D(D^T D)^{-1} D^T - \frac{1}{N} J_{N \times N})\, Y$$

$$SSE = Y^T (I_N - D(D^T D)^{-1} D^T)\, Y$$

제 3 장 스코어 방법과 HTR 방법 비교

반수체와 형질간의 관련성을 검정하는 방법으로 환자군과 정상군에서 반수체 빈도의 차이를 비교하는 분할표 χ^2 검정은 질병의 유무를 나타내는 이분형 자료에만 적용할 수 있고 반수체 이외의 변수를 고려하지 못한다는 제한점이 존재한다. 질적형질 뿐 아니라 양적형질에서도 반수체 관련성을 검정하기 위하여 선형모형에 기초한 방법을 사용하게 되었으며, 선형모형을 사용하면 환경요인 등 비유전요인의 영향을 제어할 수 있다. 선형모형에 기초한 관련성분석 방법으로 스코어 방법과 HTR 방법이 있는데, 이 두 방법에서는 EM 알고리즘을 사용하여 반수체 빈도를 추정하고 회귀모형에서 반수체 요인에 대한 회귀계수 β 가 의미 있는지를 검정한다.

스코어 방법에서는 먼저 형질의 분포에 따라 일반화 선형 모형을 세우고, 스코어 통계량을 구하게 된다. 일반화 선형 모형에서는 유전요인과 환경요인을 동시에 고려하며 환경요인에 대한 영향을 배제시키고 난 후 유전요인에 대한 스코어 통계량을 구할 수 있다. 스코어 방법은 유전형에 대한 반수체 조합(haplotype combination)의 사후확률을 계산하고, 사후 기대값을 구하여 그것을 가중치로 사용한 선형모형을 세워 유전요인의 영향이 유의한지를 검정한다. 스코어 통계량은 χ^2 분포를 따르며 검정을 통해 반수체와 형질간의 관련성이 없다는 귀무가설을 검정한다.

HTR 방법은 가중치를 준 회귀모형을 쓰는 대신에 반수체가 유전되는 사후 기대수를 구하여 하나의 독립변수로 하고, 양적 형질을 종속변수로 하여 단순 회귀분석을 시행한다. 기존의 선형모형을 통한 반수체와 형질간의 관련성분석이 표본크기가 두 배수가 된 분산분석 방법으로 이루어졌음을 보완하여 표본크기를 유지한 분산분석 방법을 사용하며 F 검정을 하게 된다.

스코어 방법과 HTR 방법은 형질의 형태에 제한받지 않고 모두 반수체 전체에 대한 관련성 분석 결과와 함께 각 반수체에 대한 관련성 분석 결과도 보여준다는

장점이 있다. 또 window sliding 방법을 통해서 반수체와 형질의 관련성을 검정하는 가장 적절한 반수체크기(window size)를 볼 수 있다.

HTR 방법과 비교해서 스코어 방법의 큰 장점은 유전요인 뿐 아니라 환경요인까지 생각할 수 있다는 점이다. 또 HTR 방법에서는 양적형질이 정규분포를 따라야 하는데, 스코어 방법에서는 형질의 분포가 가정된다면 적절한 분포에 대한 일반화 선형 모형을 세울 수 있으며 스코어 통계량도 구할 수가 있다.

이 논문에서는 반수체와 양적형질의 관련성분석을 위하여 선형모형을 이용한 관련성 검정 방법인 스코어 방법과 HTR 방법을 비교해 보았다. 실제 자료를 이용하여 두 방법을 적용하였는데 스코어 방법은 프로그램 R을 사용하였고 HTR 방법은 LINUX를 사용하여 분석하였다.

제 4 장 실제 자료를 이용한 분석

4.1 실제 자료 준비

심혈관계질환 유전체연구센터(www.heartgenome.org)에서 수집된 유전체자료(CGC자료)를 이용하여 스코어 방법과 HTR 방법을 적용한 반수체와 양적형질의 관련성을 검정하였다. 유전체자료에서 혈연관계가 없는 개인을 표본으로 하였고 하나의 유전자 위에 두 개 이상의 SNP이 존재하여 반수체를 구성할 수 있는 모든 유전체자료를 사용하였다. 구성할 수 있는 반수체는 모두 12개로, 각 반수체는 두 개의 SNP으로 구성된 것에서 여섯 개의 SNP으로 구성된 것까지 있다. 전체 표본은 3344명이고 개인마다 분석되어 있는 유전체 자료가 다르므로 반수체마다 분석되는 표본수에 차이가 있다. 반수체와 반수체를 구성하는 SNP수에 대한 자료는 아래의 표와 같다.

표 2. 심혈관유전체자료(CGC자료)에서 유전체자료 현황

Gene name		SNP수	표본수
Angiotensin I converting enzyme	ACE	6	1506
Angiotensinogen	AGT	4	1591
Arachidonate 5-lipoxygenase-activating protein	ALOX5AP	3	264
Adiponectin gene	APM	2	271
Apolipoprotein A1	APOA1	2	775
Apolipoprotein A5	APOA5	5	134
Cholesteryl ester transfer protein	CETP	3	1392
Endothelial adhesion molecule 1	ESEL	3	579
Hepatic lipase	LIPC	2	811
Microsomal triglyceride transfer protein	MTP	2	415
Toll-like receptor 4	TLR	3	271
Tumor protein, translationally-controlled 1	TPT	3	685

반수체를 추정하기에 앞서 반수체를 구성할 단일 표식유전자들이 서로 연관불균형(linkage disequilibrium) 관계에 놓여있는지 검정하는데, 유전체자료 중 LIPC를 구성하는 단일 표식유전자들이 연관균형관계(linkage equilibrium)에 놓여있다는 귀무가설을 기각하지 못했다(p=0.532). 나머지 반수체들을 구성하는 단일 표식유전자들은 연관불균형관계에 놓여있는 것으로 확인되어 LIPC를 제외한 나머지 반수체에 대한 관련성 분석을 시행하였다.

분석의 대상이 되는 양적형질은 수집된 자료의 혈액검사 결과로, 총 콜레스테롤(total cholesterol : TCHOL), 중성지방(triglyceride : TG), 저밀도 지단백(low density lipoprotein : LDL-C), 고밀도 지단백(high density lipoprotein : HDL-C), 혈당(glucose), 인슐린(insulin), 인슐린저항성지표(Homa index : HOMA), 크리아티닌(creatinine), apolipoprotein AI(APOAI), apolipoprotein B(APOB), 요산(uric acid)들의 수치를 사용하였다. 선형모형을 이용한 관련성 분석에서 종속변수는 정규분포가정을 하기 때문에 형질을 변환시켜서 정규분포를 따르도록 하였다. 정규변환을 위한 방법으로는 boxcox방법을 사용하였으며 모든 분석에 정규성을 따르는 양적형질을 이용하여 반수체의 관련성분석을 하였다. 표 3 에서는 정규화 시키기 위한 λ 값과 변환된 양적형질의 평균, 표준편차가 나와 있다.

표 3. 정규화시킨 양적형질의 분포

양적형질	λ	변환된 형질 평균	변환된 형질 표준편차
TCHOL	0.2	201.87	38.20
HDL-C	0.2	44.90	11.94
LDL-C	0.6	127.19	33.57
TG	-0.2	125.54	71.04
Glucose	-1.2	88.48	16.86
APOAI	0.6	132.04	27.07
APOB	0.4	90.29	24.92
Insulin	0.2	7.55	4.39
HOMA	0	1.63	1.10
Uric Acid	0.4	5.04	1.53
Creatinine	0.2	0.77	0.23

4.2 분석 결과

혈액검사 결과인 양적형질과 추정된 반수체의 관련성 검정 결과, HTR 방법이 스코어 방법보다 검정력이 높은 경향을 나타냈으며 이러한 차이는 형질과 반수체의 관련성이 유의하게 나타날수록, SNP수가 많은 반수체 일수록 커졌다. 아래 그래프들은 반수체와 양적형질의 관련성 분석 결과인 유의확률(p-value)을 -LOG10으로 취하여 나타낸 것이다(그림3-그림10). 관련성이 없다는 귀무가설을 기각하기 위한 유의수준을 0.05라 했을 때 -LOG10(0.05)의 값은 1.3이므로 -LOG10(p-value) > 1.3 일 때 반수체와 형질의 관련성이 존재한다고 할 수 있다. 우선 양적형질 APOAI와 HOMA, TG, Insulin, HDL-C, glucose 에서 유의한 관련성을 보이는 반수체가 존재했는데, HTR 방법과 스코어 방법을 적용한 결과에 차이가 있었다. HTR 방법을 적용하였을 때는 양적형질 APOAI, Insulin과 반수체 APOA5가 유의한 관련성을 갖는 것으로 나타났지만, 스코어 방법을 적용한 결과는 유의수준 0.05에서 관련성이 없다는 귀무가설을 기각하지 못했다(그림3, 그림4). 반수체 ACE에 대해서도 HTR 방법에서는 양적형질 APOAI, Insulin, TG, HOMA와 관련성이 없다는 귀무가설을 기각하지만 스코어 방법에서는 기각하지 못하였다(그림3, 그림4, 그림5, 그림7). HTR 방법과 스코어 방법 모두에서 유의한 관련성을 보인 양적형질 TG, HDL-C와 반수체 APOA5의 결과는 HTR 방법에서 더 유의한 결과를 보였다(그림5, 그림6). 양적형질 glucose에서는 반수체 CETP가 HTR 방법과 스코어 방법에서 모두 유의한 관련성이 있다는 결과를 나타냈으며, 역시 HTR 방법에서 더 유의한 결과를 보였다(그림10).

전체적인 결과는 반수체와 양적형질의 관련성 검정 방법에서 스코어 방법보다 HTR 방법이 검정력이 높은 경향을 보이며, 관련성이 유의하다는 결과가 나올 때 두 방법의 결과 차이가 분명해지고 있었다. 또 반수체를 구성하는 SNP수에 영향을 받는 것으로 보여, 2개의 SNP으로 구성된 반수체부터 6개의 SNP으로 구성된 반수체까지 순서대로 나열하여 결과를 비교하였다.

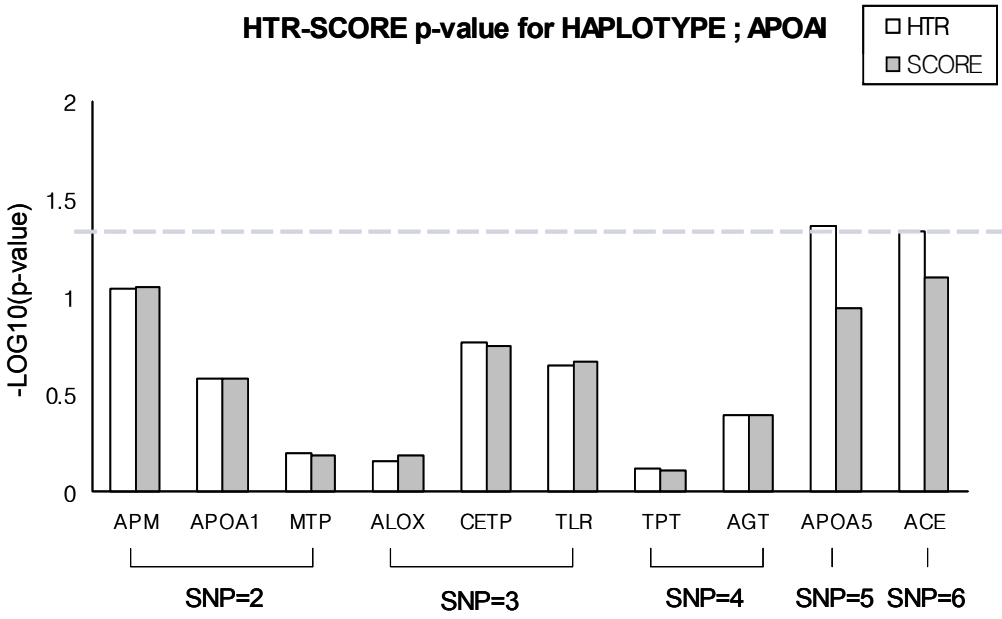

그림 3. 양적형질 APOAI에 대한 HTR-스코어방법 적용 결과

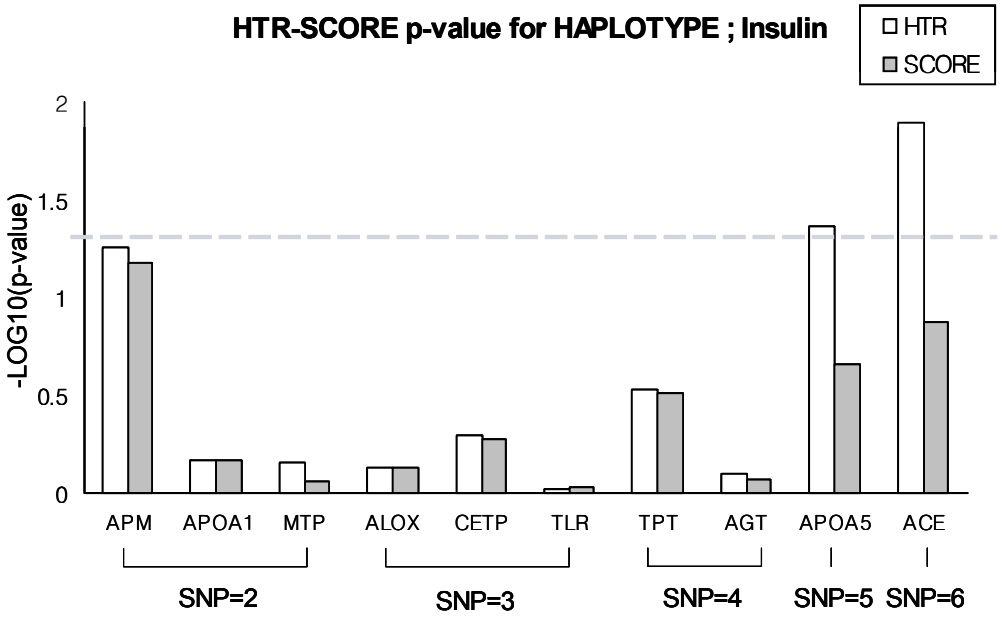

그림 4. 양적형질 Insulin에 대한 HTR-스코어방법 적용 결과

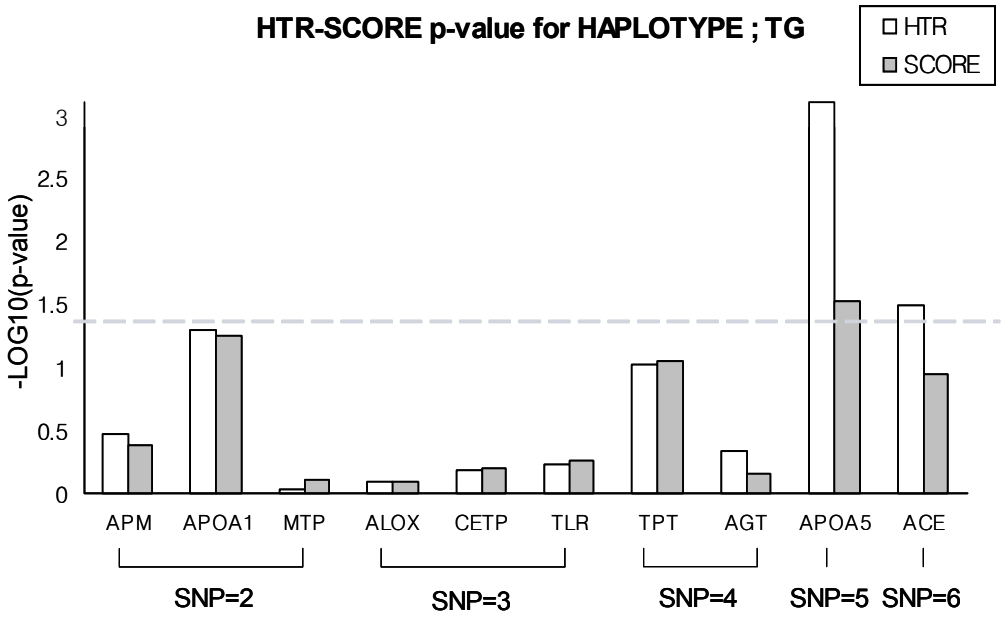

그림 5. 양적형질 TG에 대한 HTR-스코어방법 적용 결과

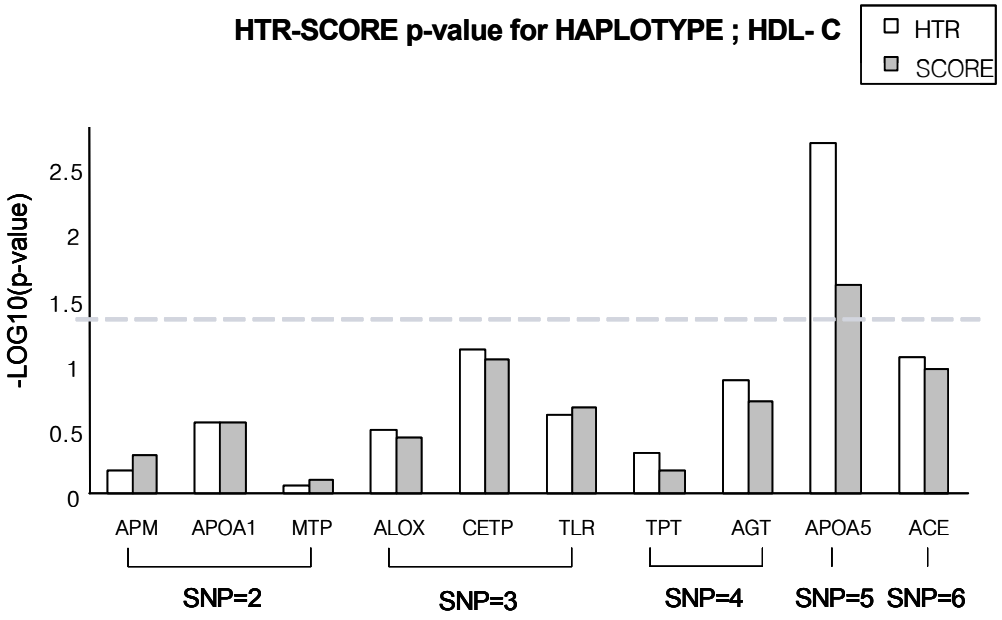

그림 6. 양적형질 HDL-C에 대한 HTR-스코어방법 적용 결과

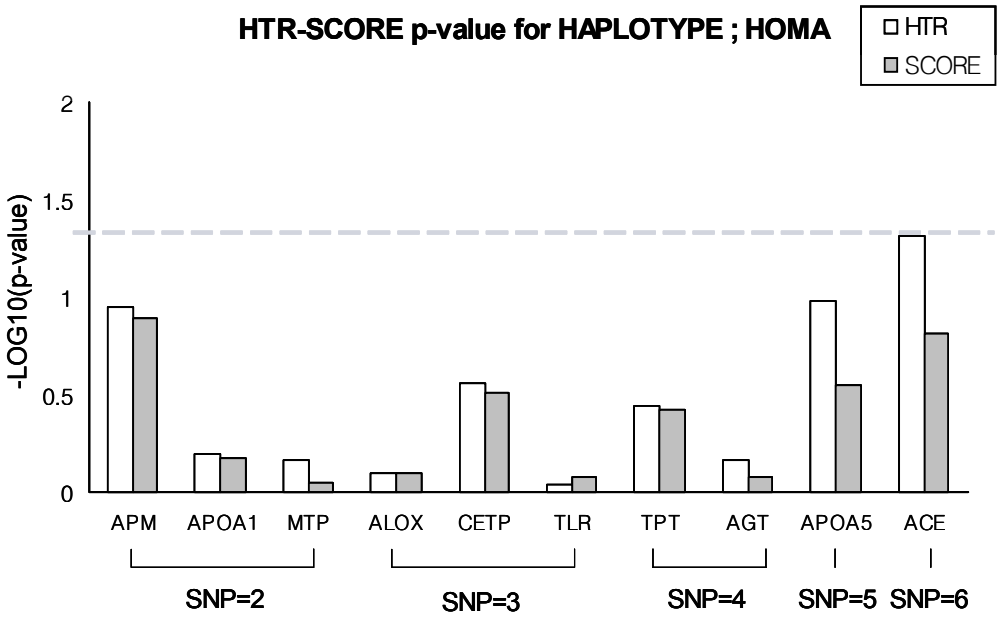

그림 7. 양적형질 HOMA에 대한 HTR-스코어방법 적용 결과

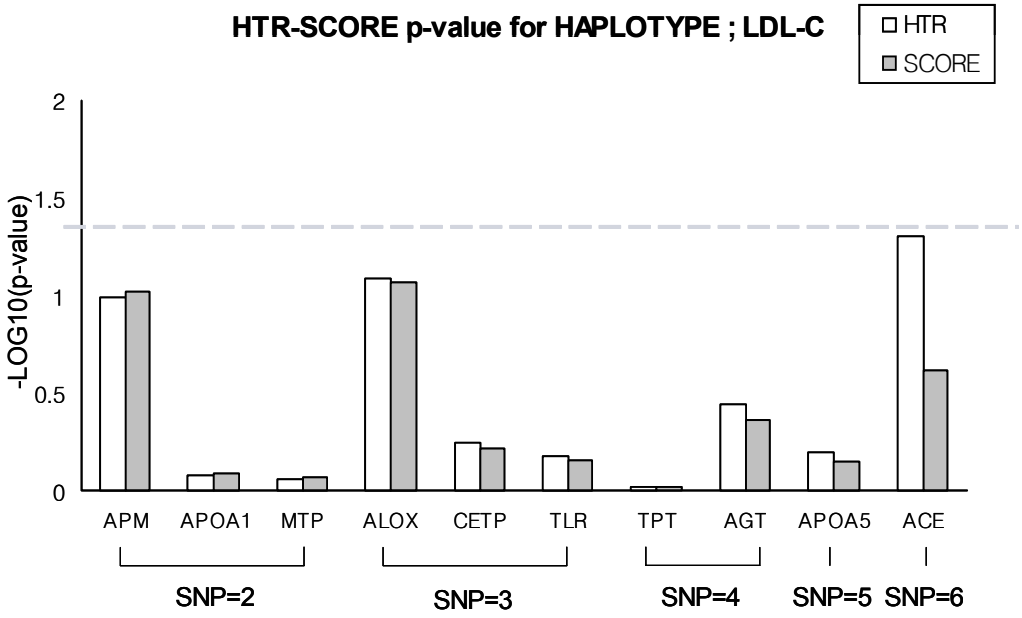

그림 8. 양적형질 LDL-C에 대한 HTR-스코어방법 적용 결과

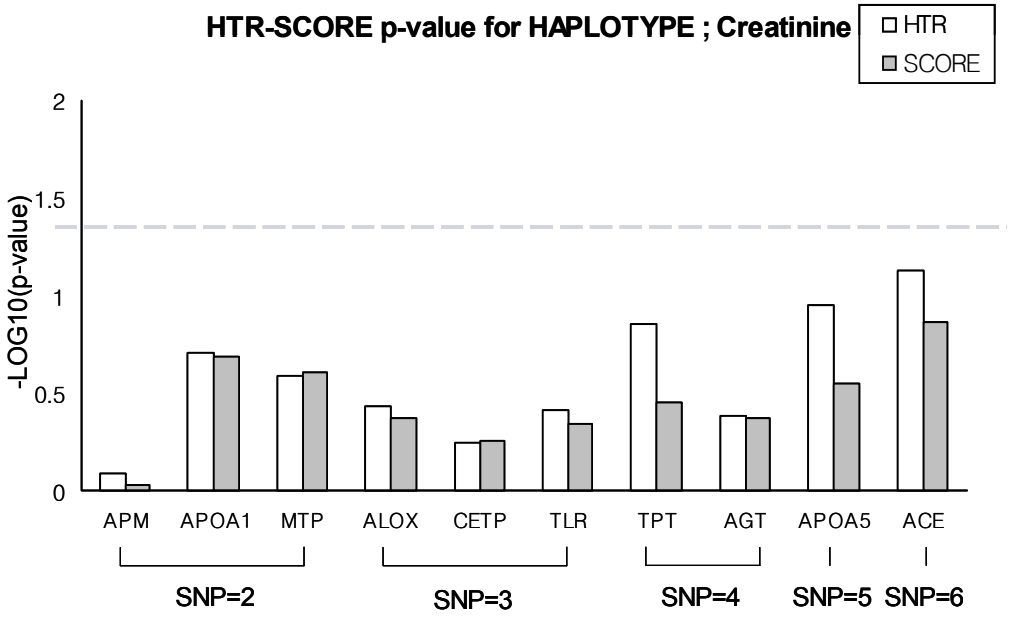

그림 9. 양적형질 Creatinine에 대한 HTR-스코어방법 적용 결과

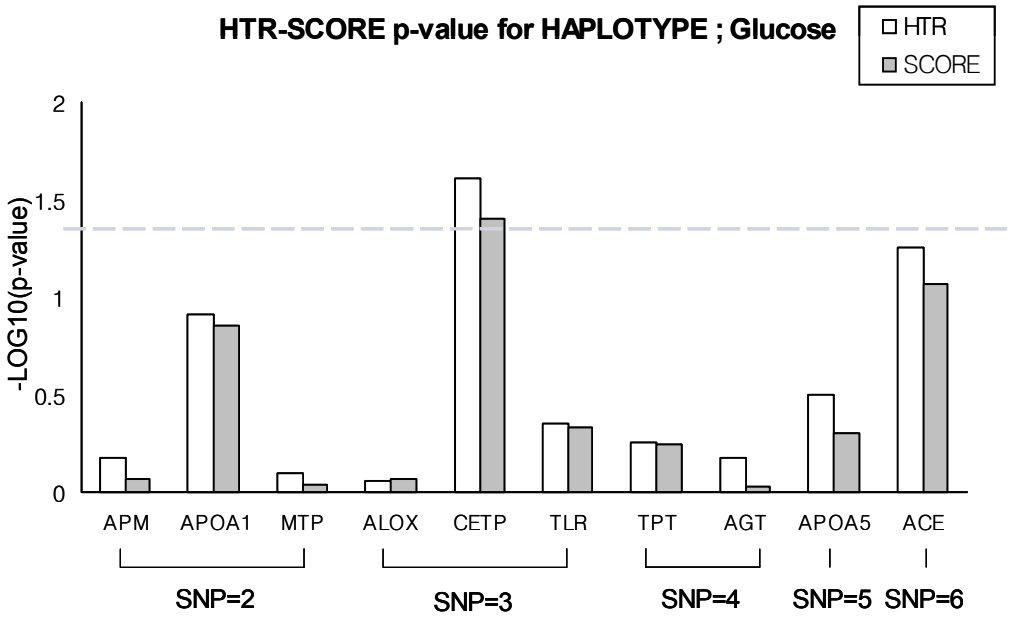

그림 10. 양적형질 Glucose에 대한 HTR-스코어방법 적용 결과

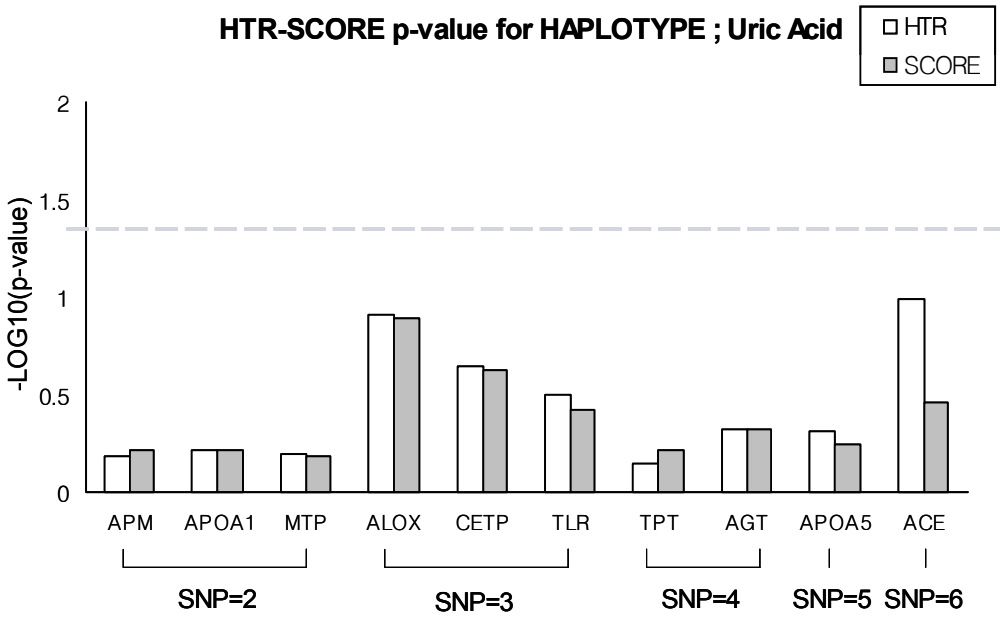

그림 11. 양적형질 Uric acid에 대한 HTR-스코어방법 적용 결과

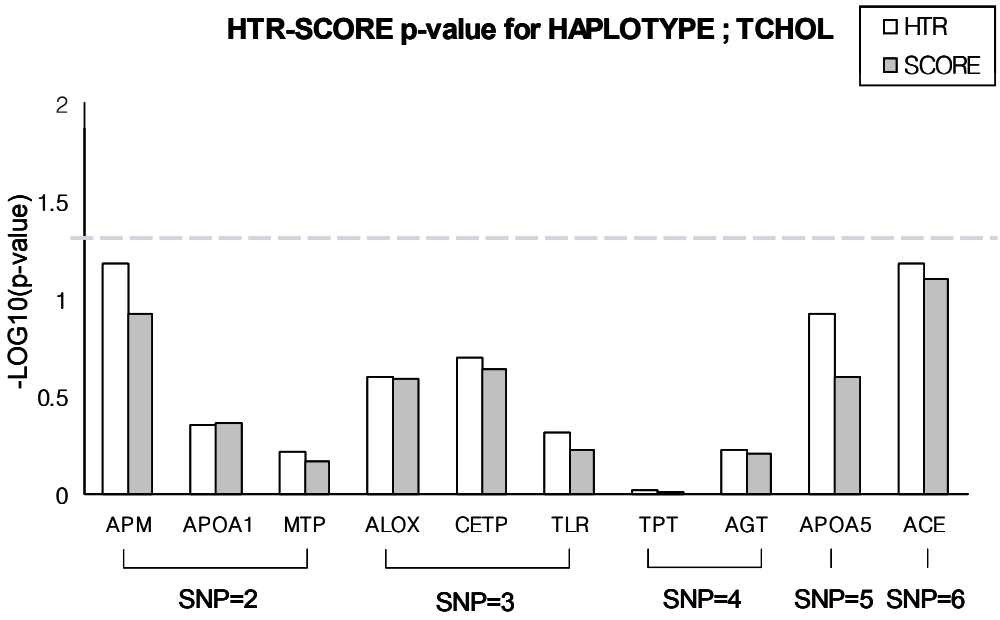

그림 12. 양적형질 TCHOL에 대한 HTR-스코어방법 적용 결과

대부분의 형질에서 반수체 ACE와 APOA5에 대한 관련성 검정 결과의 차이가 나타났는데, ACE와 APOA5는 유전체 자료에서 가장 많은 SNP를 가지고 있는 반수체로, ACE는 A-240T, C-93T, 14094, G14480C, T849S, A22982G의 6개 SNP으로 구성된 반수체이며, APOA5는 C-1399T, T-1131C, G-1029A, G-12A, 1259T/C의 5개 SNP으로 구성되어 있다.

반수체를 구성하는 SNP의 수가 HTR 방법과 스코어 방법을 사용한 관련성 분석 결과의 차이에 영향을 미치는지 자세하게 보기 위하여 반수체 ACE와 APOA5에서 window-sliding 방법을 이용하여 양적형질 APOAI와 HOMA, HDL-C 에 대한 관련성 분석을 하였다. ACE는 6개의 SNP으로 구성되었기 때문에 SNP을 두 개씩 묶어서 분석했을 때와, 세 개씩, 네 개씩 , 다섯 개씩 묶어서 관련성을 분석하였다. APOA5는 다섯 개의 SNP이므로 두 개씩, 세 개씩, 네 개씩 SNP을 묶어서 분석하였다. window-sliding방법을 사용하여 관련성분석을 하면 적합한 window 크기, 즉 적합한 반수체 크기를 알 수 있다. 분석결과는 아래 그림과 같다.

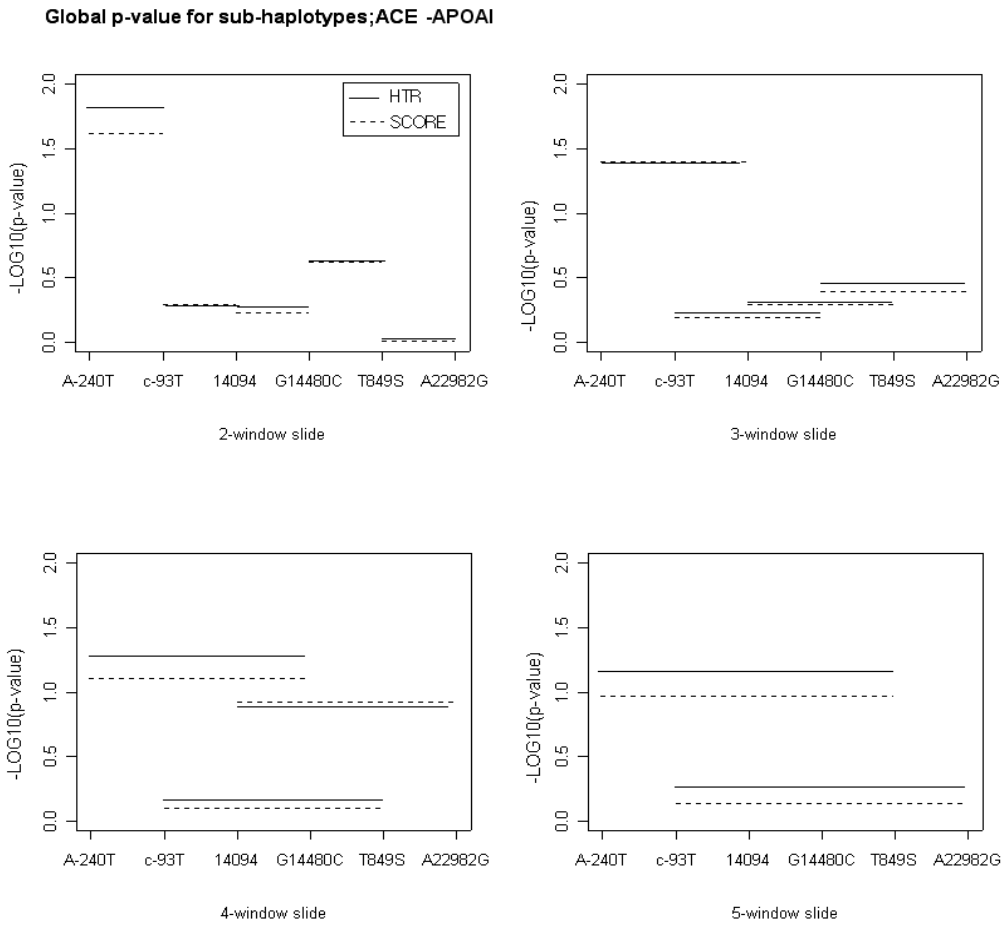

그림 13. window-slide를 이용한 형질 APOAI와 반수체 ACE의 관련성분석

반수체 ACE와 양적형질 APOAI에 대한 window-sliding방법을 사용한 결과이다. 반수체 ACE를 구성하고 있는 6개의 SNP에서 먼저 두 개의 SNP으로 묶은 반수체들의 관련성분석을 하고 차례로 세 개의 SNP, 네 개의 SNP, 다섯 개의 SNP으로 묶은 반수체로 양적형질 APOAI와 관련성분석을 하였다. 실선이 HTR 검정 결과이며 점선이 스코어 방법의 결과이다. 2-window 에서 5-window로 갈수록 HTR 방법과 스코어검정 방법의 유의확률이 차이가 커지고 있다.

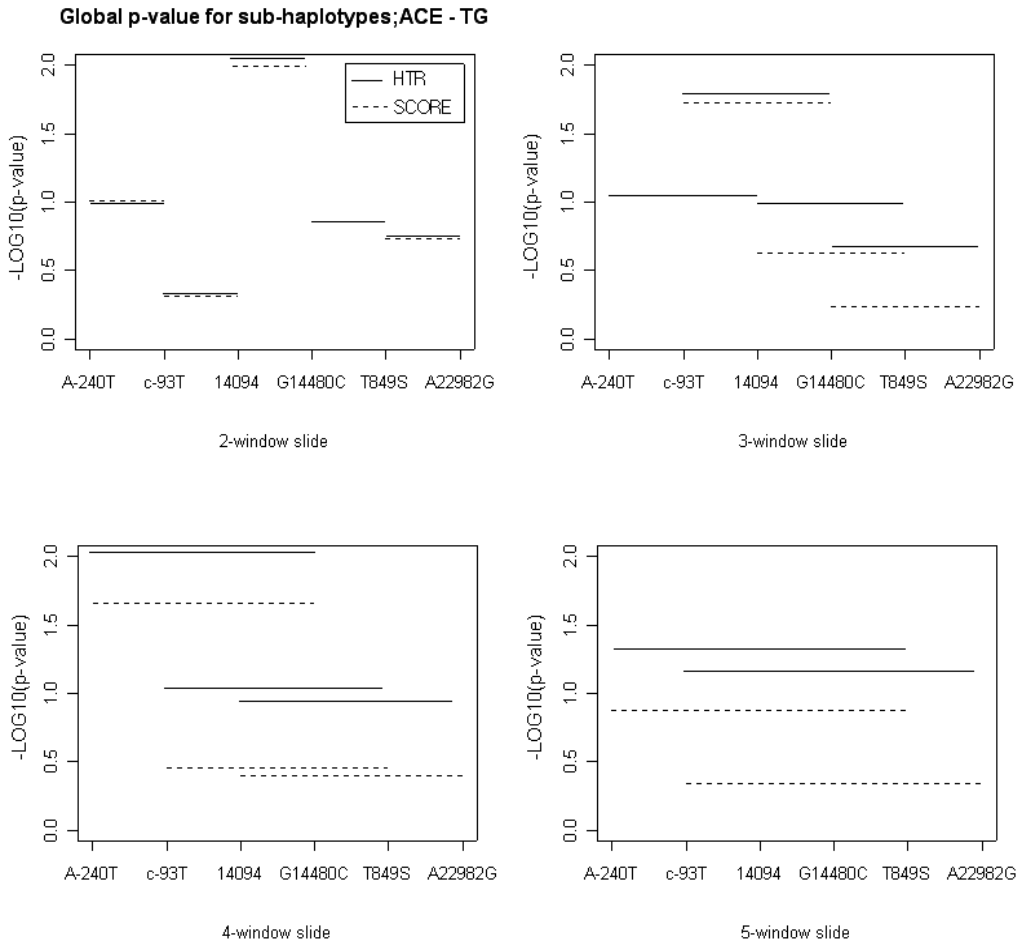

그림 14. window-slide를 이용한 형질 TG와 반수체 ACE의 관련성분석

반수체 ACE와 양적형질 TG에 대한 window-sliding방법을 사용한 결과이다. 관련성 결과는 14094, G14480C 두 개의 SNP으로 만들어진 반수체일 때 TG와 가장 높은 관련성을 갖는다. 2-window에서 HTR 방법과 스코어검정 방법의 결과 차이는 거의 없지만 3-window 이상에서는 유의확률이 차이가 나타나며 window크기가 커질수록 두 검정결과의 유의확률의 차이도 확연히 커지고 있다.

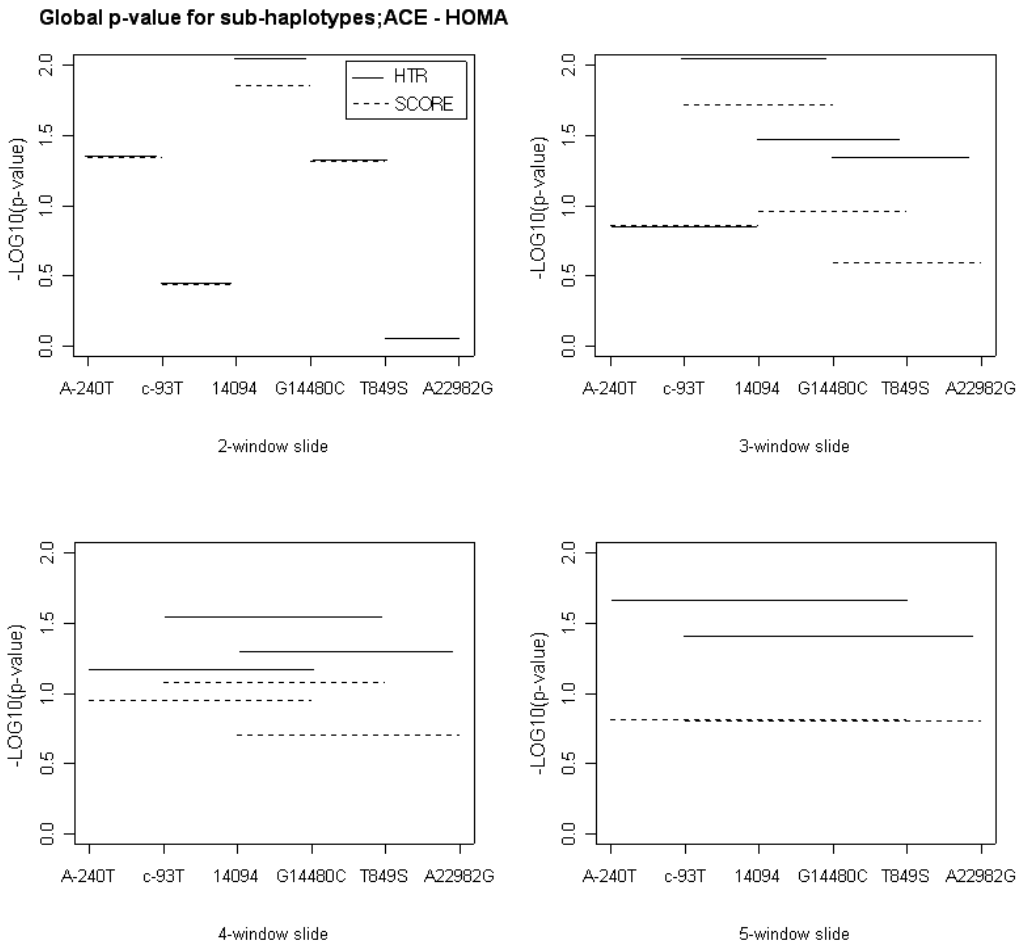

그림 15. window-slide를 이용한 형질 HOMA와 반수체 ACE의 관련성분석

반수체 ACE와 양적형질 HOMA에 대한 window-sliding방법을 사용한 관련성 분석 결과이다. 2-window에서는 14094, G14480C로 구성된 반수체일 때 높은 관련성을 갖으며 그 때에만 HTR 방법과 스코어 방법의 결과 차이를 보인다. 4-window에서 HTR 방법을 이용한 결과는 스코어 방법을 이용한 결과와 모든 SNP 범위에서 차이를 보이며 , 5-window에서 유의확률의 차이가 가장 커진다.

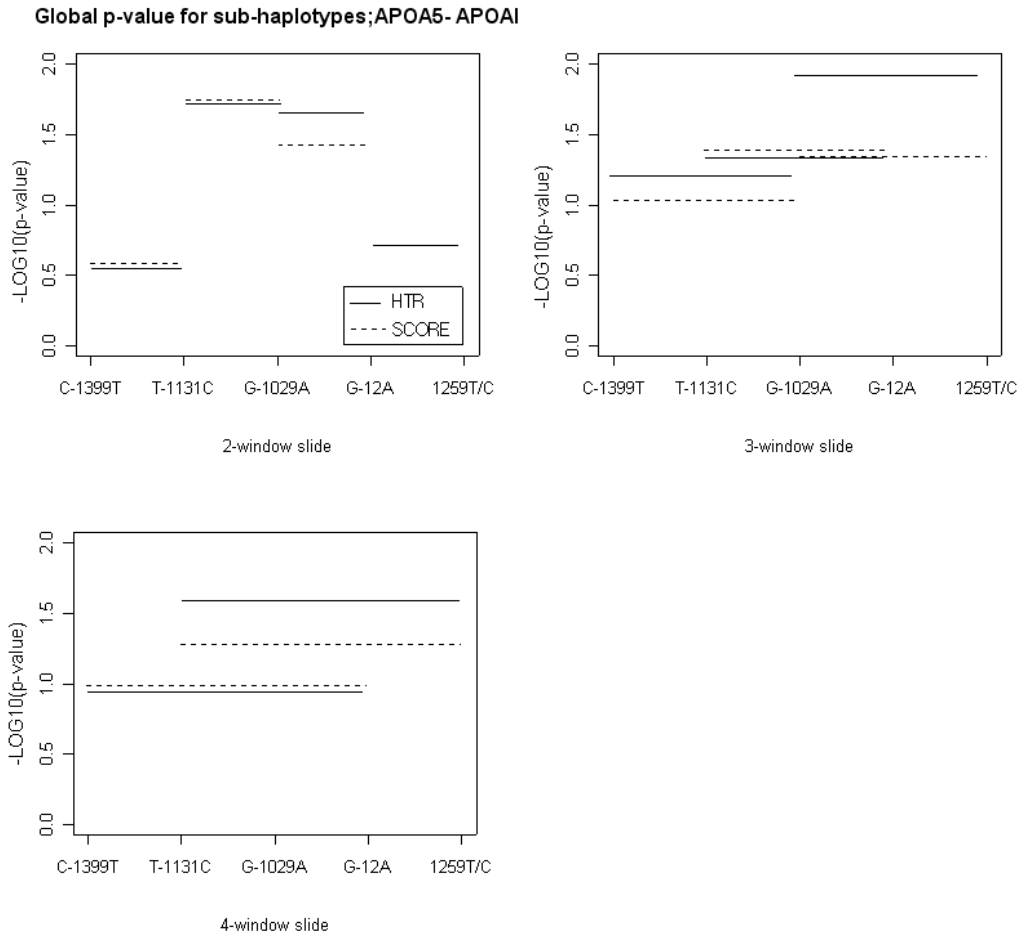

그림 16. window-slide를 이용한 형질 APOAI와 반수체 APOA5의 관련성분석

반수체 APOA5와 양적형질 APOAI에 대한 window-sliding방법을 사용한 결과이다. 2-window에서 T-1131C, G-1029A로 구성된 반수체일 때 가장 유의한 관련성을 갖는다. HTR 결과에서는 2-window에서 G-1029A, G-12A와 G-12A, 1259T/C에서의 관련성보다 3-window에서 G-1029A, G-12A, 1259T/C를 반수체로 묶었을 때 더 유의한 관련성을 갖는 것으로 보여준다. 4-window 에서도 스코어 방법 결과와 HTR 방법 결과의 차이가 있다.

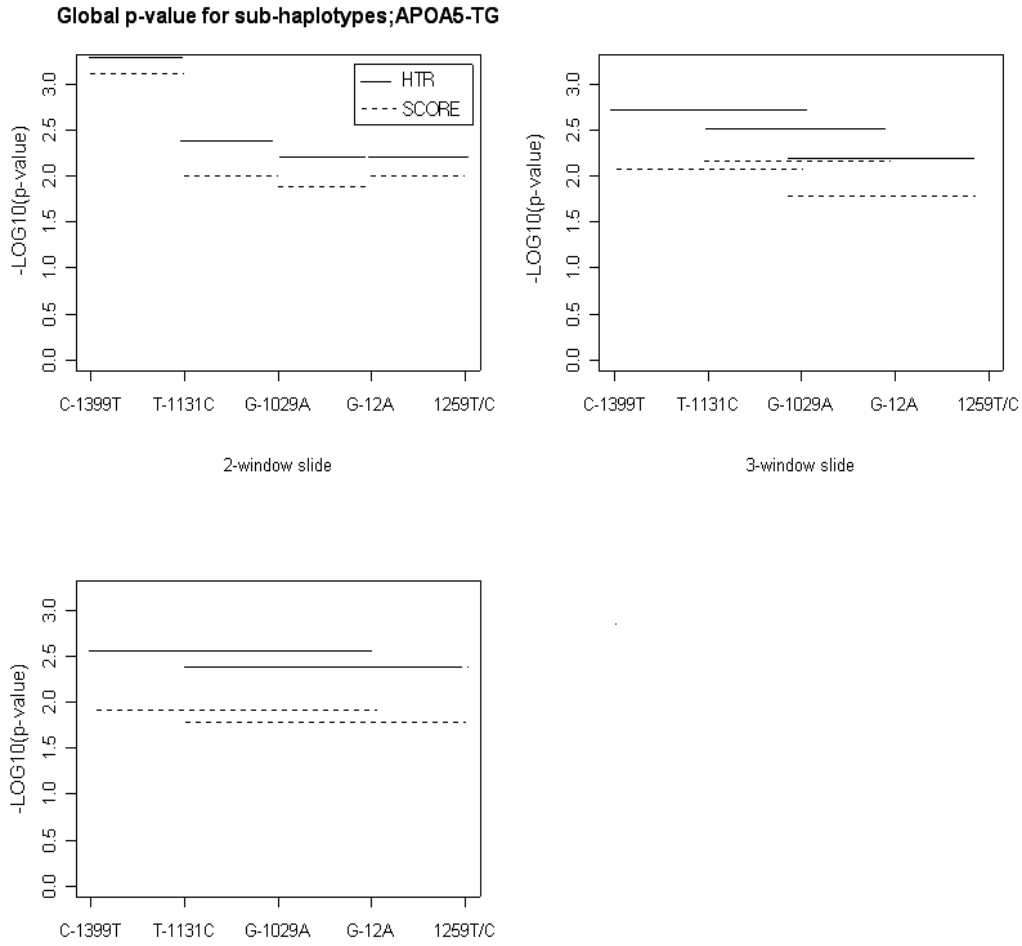

그림 17. window-slide를 이용한 형질 TG와 반수체 APOA5의 관련성분석

반수체 APOA5와 양적형질 TG에 대하여 window-sliding방법을 사용한 반수체 관련성 분석 결과이다. 우선 TG는 APOA5 전체 SNP에서 관련성이 높게 나타났다. 2-window 에서 C-1399T, T-1131C 로 구성된 반수체일 때 가장 유의한 관련성을 갖으며, 여기에서는 2-window에서부터 HTR 방법과 스코어 방법에서 유의확률의 차이가 나타났다. 4-window slide에서는 두 방법 결과 유의확률의 차이가 가장 크게 보여진다.

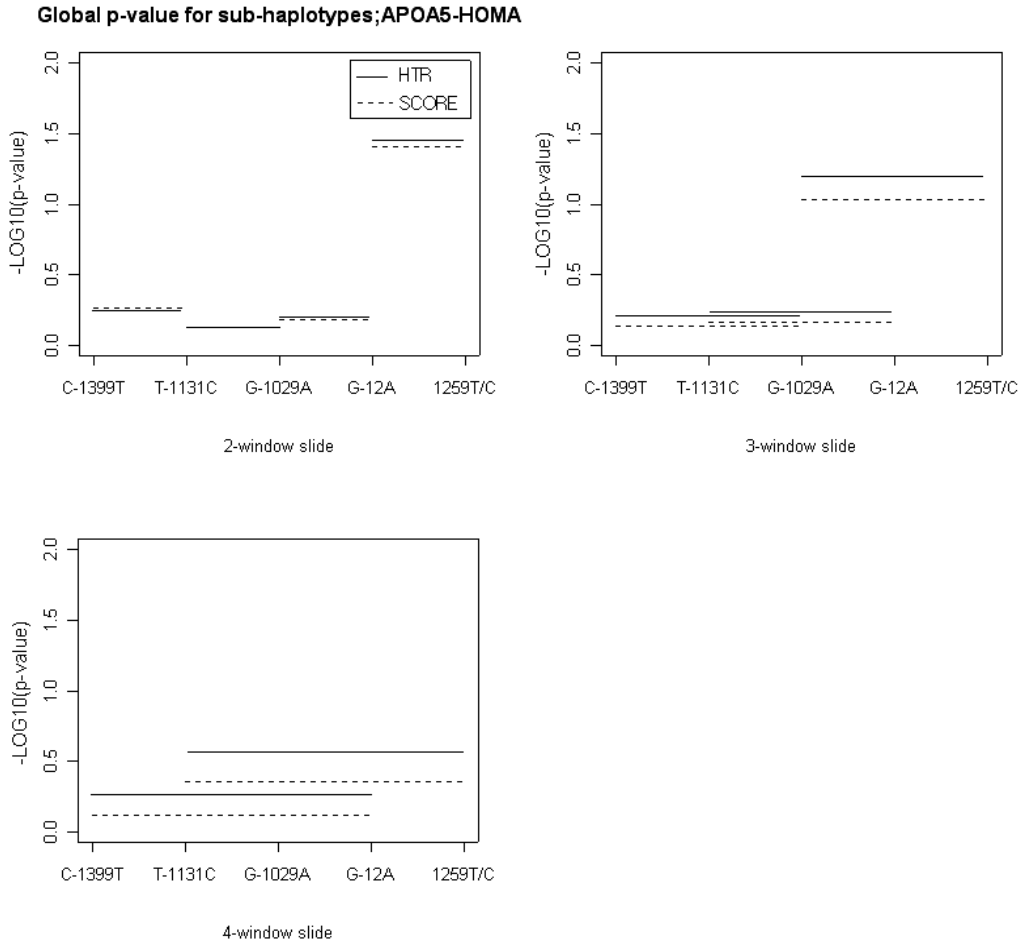

그림 18. window-slide를 이용한 형질 HOMA와 반수체 APOA5의 관련성분석

반수체 APOA5와 양적형질 HOMA에 대해 window-sliding방법을 사용한 반수체관련성분석 결과이다. APOA5와 양적형질 HOMA는 2-window에서 G-12A, 1259T/C를 제외하고는 유의한 관련성을 보이지 않으며, 두 방법 결과의 차이가 거의 나타나지 않았다. 3-window slide에서는 G-1029A, G-12A, 1259T/C로 이루어진 부분에서 두 방법의 차이가 보였다. 4-window slide에서는 관련성이 없음에도 HTR방법과 스코어방법과의 차이가 크다.

제 5 장 토의 및 결론

지금까지 양적형질과 반수체의 관련성을 검정하는 선형 모형을 기반으로 한 방법 중 스코어 방법과 HTR 방법을 실제 자료를 이용하여 적용하고 비교하였다. 두 방법 모두 기본 개념은 EM 알고리즘으로 반수체 빈도를 추정하고 선형모형에 기초하여 반수체와 양적형질간의 관련성을 검정하는 것이다.

본 논문에서는 실제 자료에서 반수체를 구성할 수 있는 모든 유전체 자료와 혈액검사 결과 자료를 이용하여 양적형질과 반수체의 관련성 분석을 하고 그 결과를 비교하였다. 분석 결과 HTR 방법이 스코어 방법보다 검정력이 높은 경향을 보이고 있었다. 이러한 경향은 반수체와 형질간의 유의한 관련성이 있을 때 더 많은 차이를 보이고 있었으며, 반수체를 구성하는 SNP(locus)이 많아질수록 두 검정결과의 차이는 커지는 양상을 보였다. 이 결과는 window-sliding 방법을 통해서 더 자세히 알아볼 수 있었다. window-sliding 방법에서는 반수체를 구성하는 SNP 수가 5개 이상인 반수체 APOA5와 ACE 반수체를 사용하였다. APOA5는 SNP을 2개부터 4개까지 ACE는 SNP을 2개부터 5개까지 하나씩 늘려가면서 양적형질의 관련성을 분석 하였다. 그 결과 하나의 반수체 안에서도 반수체를 구성하는 SNP 수가 늘어남에 따라서 HTR 방법의 결과가 스코어 방법 결과보다 유의확률이 좀 더 낮은 경향을 보이고 있었다. 이러한 경향은 HTR 방법이 반수체가 유전되는 사후 기대수 만을 독립변수로 사용하여 양적형질과 단순회귀분석으로 검정하는 것으로서 스코어 방법보다 직접적이기 때문에, SNP이 많아질수록 HTR 방법의 검정력이 더 커지는 것으로 생각된다.

본 논문의 분석 결과에서 SNP수가 큰 반수체에서 관련성을 검정 할 때에는 HTR 방법을 사용하는 것이 더 유용하다는 점을 제시하고 있다. 여기서는 반수체와 양적형질 자료만을 이용하여 관련성을 검정하였는데, 질병과 관련된 복잡한 형질들은 유전적 요인과 환경요인의 영향을 동시에 받기 때문에 환경요인을 고려했을 때의 스코어 방법과 HTR 방법의 차이를 알아보는 것이 필요하다. 스코어 방

법은 환경적 요인의 영향과 유전요인의 영향을 동시에 고려하여 유전요인에 대한 관련성 결과를 제시하지만, HTR 방법에서는 오직 유전요인에 대한 영향만을 고려하므로 환경요인에 대한 영향을 고려하는 방법에 대한 연구가 진행되어야 할 것이다. 또한 유전요인과 환경요인을 생각할 뿐 아니라 유전요인과 유전요인의 교호작용, 유전요인과 환경요인의 교호작용까지 고려하는 방법이 있어야 할 것이다. 반수체는 반수체를 구성하고 있는 SNP수가 많아질수록 추정해야 하는 SNP 조합의 수가 많아지기 때문에 검정력이 떨어질 수 있다. 따라서 반수체수의 증가로 자유도가 커질 때 검정력이 떨어지는 경우를 보완해야 하는 점에 대해서도 논의되어야 할 것이다.

참 고 문 헌

김민지. (2003) 형제자료에 대한 양적형질의 유전자 관련성 분석방법의 비교. 연세대학교 석사학위논문.

박찬미. (2002) 양적형질의 유전자 연관성 분석을 위한 개선된 헤이즈만엘스톤방법과 분산성분방법의 비교. 연세대학교 석사학위논문.

송기준. (2003) 양적형질 유전자의 연관 및 관련성에 대한 동시적 분석. 연세대학교 박사학위논문.

Akey, J., Jin, L., Moniao, X. (2001) Haplotypes vs single marker linkage disequilibrium tests : what do we gain?, *Eur J Hum* Genet 68: 191-197.

Becker, T., Knapp, M. (2004) Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet Epidemiol* 27: 21-32.

Cardon, L. R. (2003) Using haplotype blocks to map human complex trait loci. *Trends in Genet* 19: 135-140.

Chapman, J. M., Cooper, J. D., Todd, J. A., Clayton, D. G. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 56: 18-31.

Chiano, M. N., Clayton, D. G. (1998) Fine genetic mapping using haplotype analysis and the missing data problem. *Ann Hum Genet* 62: 55-60.

Clark, A. G. (1990) Inference of haplotypes from PCR-amplified samples of diploid population. *Mol Biol Evol* 7: 111-122.

Cordell, H. J., Clayton, D. G. (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphism within a gene using case/control or family data: application to HLA in the 1 diabetes. *Am J Hum Genet* 702: 124-141.

Daly, M. J. (2001) High-resolution haplotype structure in the human genome. *Nature Genet* 29: 229-232.

Epstein, M. P., Satten, G. A. (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 73: 1316-1329.

Fallin et al. (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variateion and Alzheimer's disease. *Genome Res* 1: 143-151.

Iturria, S. J., Blangero, J. (2000) An EM algorithm for obtaining maximum likelihood estimates in the multi-phenotype variance components linkage model. *Ann Hum Genet* 64: 349-362.

Judson, R., Stephens, J. C. (2001) Notes from the SNP vs haplotype front. *Pharmacogenomics* 2: 7-10.

Keavney et al. A. (1998) Measured haplotype analysis of the angiotensin-1 converting enzyme gene. *Hum Mol Genet* 1: 1745-1751.

Lin, D. Y. (2004) Haplotype-based association analysis in cohort studies of unrelated individuals. *Genet Epidemiol* 26: 255-264.

Mano, S., Yasuda, Y. (2004) Notes on the maximum likelihood estimation of haplotype frequencies. *Ann Hum Genet* 68: 257-264.

Rohde, K., Furst, R. (2003) Association of genetic traits to estimated haplotypes from SNP genotypes using EM algorithm and Markiv chain monte carlo technique. *Hum Hered* 56: 41-47.

Richard, W. M. (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemio*l 23: 221-233.

Schaid et al. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 70: 425-434.

Shibara et al.. (2004) Simultaneous estimation of haplotype frequencis and quantitative trait prameters: Applications to the test of association between phenotype and diplotype configuration. *Genetics* 168: 525-539.

Stephens, M., Nicholas, J. S. (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68: 978-989.

Thomas et al.. (2003) Bayesian spatial modeling of haplotype association. *Hum Hered* 56: 32-40.

Tianhua, N., Zhaohui S. Q. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 70: 157-169.

Tregouet et al. (2004) A new algorithm for haplotype-based association analysis: the stochastic-EM algorithm. *Ann Hum Genet* 68: 165-177.

Wallenstein, S., Hodge, S. E., Weston, A. (1998) Logistic regression model for analyzing extended haplotype data. *Genet Epidemiol* 15: 173-181.

Zaykin, D. V. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 53: 79-91.

Zhao, H., Pfeiffer, R., Gail, M. H. (2003) Haplotype analysis in population genetics and association studies. *Pharmacogenomics* 4: 171-178.

Zhao, J. H., Curtis, D., Sham, P. C., (1998) Model-free analysis and permutation tests for allelic associations. *Hum Hered* 50: 133-139.

Zhao, L. P., Li, S. S., Khalid, N. (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am J Hum Genet* 72: 1231-1250.

Zhu, X., Yan, D. (2003) Linkage disequilibrium and hgaplotype diversity in the genes of the Renin-Angiotensin system: findings from the family blood pressure program. *Genome Res* 13: 173-181.

ABSTRACT

Comparison of Regression-based Methods for Haplotype Association Analysis

Lee, Eun Hye
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

In human genetic study, exploring the associations between genes and disease phenotypes is an important step toward the discovery of genes that influences complex human diseases. Single nucleotide polymorphisms (SNPs) are currently being explored for use as genetic markers in association studies of complex disease. Since multiple SNPs in a region are likely in linkage disequilibrium, it has been suggested that methods which use the information at several SNPs at a time, along the haplotype, will be better for finding disease-predisposing genes through association studies.

A popular method of testing association between haplotypes and traits is comparing the haplotype frequences between the cases and controls. Using method based on regression model, it would possible to test the statistical association between haplotypes and a wide variety of traits, including binary, ordinal, and quantitative traits and adjust for non-genetic covariates.

In this thesis, we compared the Score test with the HTR both based on

regression model for testing association between haplotype and quantitative traits. In order to compare the results of two methods, we used Cardiovascular genomic center data containing 12 haplotypes data and blood test results as quantitative traits. The results of the association test showed that HTR has higher power than Score test when there was significant association. Difference of HTR results and Score test results had a tendency to increase as the number of SNP in the haplotype increase.

Key words : Haplotype, Association analysis, Score test, HTR, qauntitative traits.