응용통계연구 (2010)
23(1), 95–111

집단 및 가족기반연구에서의 유전적 연관성 분석 고찰: 방법론과 소프트웨어

이효정[1] · 김민지[2] · 박미라[3]

[1]고려대학교 통계학과, [2]삼성생명과학연구소 통계지원팀, [3]을지대학교 예방의학교실

(2009년 9월 접수, 2009년 12월 채택)

요 약

최근 단일염기다형성 및 일배체형을 이용한 질병-유전자간 연관성연구가 많이 진행되고 있으며, 이를 위한 다양한 분석방법과 분석도구가 개발되고 있다. 그러나 통합 소프트웨어는 충분히 확립되지 못하였으며, 각 소프트웨어가 제공하는 분석방법 및 양식에 차이가 많아 연구자가 적절한 것을 선택하기가 쉽지 않다. 본고에서는 유전적 연관성연구를 사전분석단계, 집단기반연구방법, 가족기반연구방법으로 나누어 각각의 목적에 따른 분석방법을 고찰하고, 이의 분석을 위한 주요 소프트웨어로서 FBAT, SAS/Genetics, SAGE, R의 지원내용과 방법을 비교하였다.

주요용어: 유전적 연관성, 단일염기다형성(SNP), 일배체형(haplotype), 소프트웨어.

1. 서론

유전학연구에서 대립형질(allele)이나 유전자형(genotype), 일배체형(haplotype)에 기초한 유전자와 질병간의 연관성 분석(association analysis)은 매우 중요한 부분 중 하나이다. 유전표지자로서 과거에는 초위성체(microsatellite) 마커가 주로 사용되었으나, 최근에는 단일염기다형성(Single Nucleotide Polymorphism; SNP) 마커가 많이 사용되고 있다. SNP는 염기가 개인마다 달라질 수 있는 DNA 상에 위치한 유전변이로서, 대부분 두 개의 대립형질을 가지며, 발병원인이나 치료제에 대한 반응 등 개인적인 차이를 가져오는 원인이 될 수 있다. 따라서 이를 이용하여 각 개인별로 특정 질병의 예측과 진단 및 예방을 하고자 하는 연구에 많은 관심이 집중되고 있다.

질병과의 유전적 연관성을 분석하기 위한 방법으로 연쇄불균형(linkage disequilibrium; LD)이나 사례-대조 분석, 가계(pedigree) 분석, 부모자(trio) 분석 등에서 여러 가지 척도와 방법이 사용되고 있다 (Laird와 Lange, 2008; Saito 등, 2006; Zhao, 2000). 또한 여러 연구자에 의해 개개의 방법론을 적용하기 위한 프로그램이 개발되고 있다(http://linkage.rockefeller.edu/soft/). Laird와 Lange (2008)은 가족기반연관성 분석에 대한 소프트웨어(S/W)로서 APL, PDT, FBAT, PBAT, Golden Helix, QTDT의 특성을 간략히 비교한 바 있다. 그러나 아직까지 통합분석을 제공하는 S/W는 충분히 확립되지 못하였으며, 각 S/W가 제공하는 분석방법에도 차이가 있다.

본 고에서는 유전적 연관성 분석을 집단기반연구(population-based study)와 가족기반연구(family-based study)로 나누어, 각각의 목적에 따른 분석방법을 소개하고, 대표적인 통합형 S/W를 비교하였

이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2007-531-C00018).

[3]교신저자: (301-832) 대전시 중구 용두동 143-5, 을지대학교 의과대학 예방의학교실, 부교수.
E-mail: mira@eulji.ac.kr

다. 2절에서 주요 연관성 분석을 데이터의 특성별로 분류하여 사전분석방법과 함께 소개하고, 3절에서 유전적 연관성 검정을 위한 주요 패키지프로그램으로서 FBAT, SAS/Genetics, SAGE, R에서 제공하는 관련 프로그램을 조사하여 각 프로그램별 특성을 비교하였다. 마지막으로 4절에서 비교결과를 정리하였다.

2. 유전적 연관성 분석방법

2.1. 사전분석

2.1.1. 빈도추정과 정보력지수 각 마커의 특성을 알기위해 먼저 각 마커의 모든 대립형질과 유전자형에 대한 상대빈도를 추정하게 된다. 또한 이형접합성(heterozygosity)나 다형정보력(polymorphism information content; PIC) 등의 정보력지수를 측정한다. 이형접합성은 집단에서 하디-와인버그 평형(Hardy-Weinberg equilibrium; HWE)을 가정할 때 무작위로 선정된 사람의 유전자형이 이형접합일 확률을 의미하며, 다형정보력(PIC)은 부/모/자의 유전자형이 주어졌을 때 주어진 부모에게서 아이에게 전이되는 대립형질을 구분할 수 있는 확률을 의미한다. 이형접합성이나 다형정보력이 클수록 정보력이 높은 마커라고 할 수 있다.

2.1.2. 하디-와인버그평형과 연관불균형 이상적인 집단에서 각 유전자좌(locus)에서 부모로부터 받는 두 개의 대립형질은 서로 독립이 되는데 이러한 상태를 하디-와인버그 평형이라고 한다. 하디-와인버그 평형은 자연선택이나 돌연변이, 이주, 또는 동계교배에 의해 깨지게 된다 (Sham, 1998). 또한 집단 내에 새로운 아집단(subpopulation)이 생겼거나 질병과의 연관성이 있을 때에도 발생하게 되므로 이의 검토가 유용한 단서가 될 수 있다. 실제로는 실험상의 오류로 인해 나타나는 경우가 많아서, 데이터의 질적인 평가를 위한 방편으로 사용되고 있다. 하디-와인버그평형에 대한 p-값이 매우 작으면, 해당하는 유전자좌를 제외하고 분석하게 된다. 하디-와인버그평형의 검정을 위해서는 카이제곱 적합도 검정이나 Fisher의 정확검정, Guo와 Thompson (1992)의 정확검정 등이 사용된다.

연쇄불균형은 한 유전자좌에 있는 대립유전자의 빈도가 다른 유전자좌에 있는 대립유전자의 빈도와 독립적인지 알아보는 측도이다. 두 유전자좌 A, B에서 동시에 나타나는 대립유전자쌍의 관측확률을 p_{AB}, 두 유전자좌 각각의 대립유전자비율을 각각 p_A와 p_B라 하면, 불균형 계수는 $D = p_{AB} - p_A p_B$로 정의된다. 이 값은 대립형질의 빈도에 따라 최대값이 달라지므로, 이보다는 범위가 -1에서 1사이가 되도록 표준화한 Lewontin의 D'이나 상관계수 r^2이 주로 사용된다. 연쇄균형을 이룬 상태라면 이들 척도는 0이 된다. 일반적으로 LD가 크면 두 유전자좌간 거리가 가깝다고 할 수 있고, LD가 0이면 두 유전자좌가 염색체상에서 멀리 떨어져 있다고 생각된다. 실제 연구에서 LD를 구하는 이유는 LD가 높다면 질병의 원인이 되는 SNP가 실험에 포함되지 않았더라도 포함된 SNP와의 LD를 통해서 그 효과를 찾을 수 있을 것이라는 기대 때문이다 (Balding, 2006). 한 지역에 대한 LD는 각 대립형질간의 D'쌍 또는 r^2쌍의 평균으로 표시하거나, 값을 색상으로 표현한 그림인 LD map으로 표시한다 (Barrett 등, 2005).

2.1.3. 일배체형 추정 일배체형은 하나의 염색체상에 있는 대립형질의 조합을 의미한다. 대부분의 경우 실험을 통해 유전자형만 알 수 있으므로, 일배체형은 추정을 하게 된다. 추정방법으로 Clark의 알고리즘이나 EM 알고리즘, 베이지안 방법 등이 사용된다 (Clark, 1990; Stephens 등, 2001; Zaykin 등, 2002). 일배체형의 빈도가 각 대립형질의 빈도의 곱과 같으면 LD는 0이 되므로, 두 마커가 연쇄균형에 있다고 해석할 수 있다.

2.2. 집단기반분석

사례-대조연구와 같이 친척관계가 없는 사람들을 대상으로 하는 집단기반연구는 가족기반연구에 비해서 비교적 쉽게 표본을 구할 수 있는 장점이 있다. 그러나 아집단의 형성으로 인한 의사(spurious)연관을 일으키기 쉽고, 많은 연구가 재현되지 않는다는 보고가 있다 (Hirshhorn 등, 2002). 표현형질이 질적(이진)이냐, 양적(연속)이냐에 따라 분석방법이 달라진다. Balding (2006)은 집단기반연관연구의 형태를 분류하고 주요 분석방법을 논의하였으며, Saito 등 (2006)는 유전적 연관성연구의 보고를 위한 점검표를 제시한 바 있다.

2.2.1. 이진 형질인 경우 질병의 유/무와 같은 이진 형질(binary trait)에 대한 SNP기반의 연관성분석은 대립형질의 빈도를 비교하느냐, 유전자형의 빈도를 비교하느냐에 따라 2×2, 또는 2×3 분할표의 분석이 된다. 즉, 대립형질이나 유전자형에 대한 사례-대조 연구로 보고 카이제곱검정이나 정확검정, 오즈비 등 일반적인 범주형 자료분석에 사용되는 분석을 실시할 수 있다. 대립형질의 가법효과를 알아보려 할 때에는 Armitage의 경향검정을 실시하기도 한다. 여러 개의 대립형질을 갖는 마커라면, m개의 대립형질에 대한 $2 \times m$ 분할표를 분석하거나 가능한 유전자형의 조합에 대한 분할표 분석이나 다중대립형질 경향 검정(multiallelic trend test)을 수행할 수 있다 (Slager과 Schaid, 2001). 여러 개의 마커들을 설명변수로 놓거나, 환경요인과의 교호작용, 성별이나 연령 등의 공변량을 포함하기 위해서 로지스틱회귀분석도 사용된다. 일배체형 기반의 분석에서는 k가지의 일배체형이 생성되고, 각 개인에 대한 일배체형이 추정되면 $2 \times k$ 분할표에 대한 분석을 실시할 수 있다. 또는 개인에 대한 일배체형을 할당하는 대신 환자군과 대조군에서의 일배체형 빈도를 추정할 수도 있다. 두 경우 모두 HWE의 가정이 필요하다. 빈도가 매우 낮은 일배체형이 있을 때에는 이들끼리 묶어서 하나의 범주로 간주하거나, 모두 제거하고 분석하기도 한다. 한 개체에 대해 둘 이상의 일배체형의 배정을 허용하면서 연관분석을 하는 방법도 개발되었다 (Zhao 등, 2000; Fallin과 Schrock, 2000).

2.2.2. 양적 형질인 경우 혈압, 체중 등과 같은 양적인 형질(quantitative trait)일 때는 각 SNP의 유전자형을 설명변수로 하고, 질병과 관련된 형질을 반응변수로 하는 회귀분석이나 분산분석을 적용할 수 있다. 회귀분석은 유전자형안에 특정 대립형질이 몇 개인지에 따라 0, 1, 2의 값을 주고, 분산분석은 각 유전자형을 그룹으로 간주하여 분석하게 된다.

2.3. 가족기반 연관성 분석

가족기반연구는 질병을 가진 사람과 그의 가족에 대한 유전정보를 이용하여 질병과 유전자간의 연관성을 분석하는 방법이다. 이러한 방법은 집단층화에 의한 연관성이 나타나는 오류를 피할 수 있으며, 대조군을 모집할 필요가 없고, 유전자-유전자 또는 유전자-환경간 교호작용을 직접 알 수 있다는 점에서 유용하다. Zhao (2000)는 가족기반연관성 연구에서의 다양한 방법을 정리한 바 있으며, Ewens 등 (2008)은 양적 형질일 때의 가족기반연구방법에 대해 QTDT와 FBAT 패키지에서 사용되는 방법을 중심으로 비교한 바 있다.

2.3.1. 이진 형질인 경우 이진 형질의 연관성 검정에서 대표적인 방법은 TDT(transmission/ disequilibrium test)로서, 환자와 환자의 부모에 대한 유전자형을 모두 알 수 있으며, 부모 중 한명은 이형접합일 때 적용된다. 이 방법은 연관이 있을 때 연쇄에 대한 검정방법으로 개발되었으며, 마커와 질병유

전자간에 연쇄와 연관이 모두 없다면 마커의 형질은 부모에게서 자식으로 랜덤하게 전이(transmit)된다는 전제에서 출발하였다. 이 방법은 환자에게로 전이되는 마커형질과 전이되지 않는 마커형질을 비교하여 검정하게 되며, 기본적인 검정통계량은 다음과 같다.

$$\mathrm{TDT} = \frac{(b-c)^2}{b+c},$$

여기서, b는 유전자형 12를 가진 부모에게서 환자에게로 형질 1이 전이되는 수이고, c는 형질 2가 전이되는 수로서 이 통계량이 자유도 1인 카이제곱분포를 따른다는 것을 이용하여 검정한다. 이는 다중대립형질에 대해서도 확장된 바 있다 (Spielman과 Ewens, 1996).

부모의 유전자형을 확보할 수 없을 때에는 형매(sibs)의 정보를 이용하기도 한다. Curtis (1997)의 방법은 가족내 여러 형매가 있을 때, 이들 중 유전자형이 가장 불일치되는 한 쌍의 질환이 있는 형매와 질환이 없는 형매자료를 이용하는 방법이다. 반면 Spielman과 Ewens (1998)의 S-TDT는 질환이 있는 형매와 질환이 없는 형매가 1명씩 있어야 하며, 각자의 유전자형이 동일해서는 안된다는 최소요건을 갖추어야 한다. Horvath와 Laird (1998)의 SDT는 S-TDT와 마찬가지의 최소요건이 요구되는데 가족의 크기가 더 큰 경우에도 유효하며, 동일한 유전자형을 갖지 않는 모든 형매의 정보를 이용한다. 한편 Knapp (1999)은 부모 중 하나 혹은 두 명의 유전자형을 알 수 없을 때 자식들의 유전자형을 이용하여 부모의 유전자형을 재구축하는 방법인 RC-TDT를 제시하였으며, Monks 등 (1998)는 다중대립형질에 대한 multiallelic S-TDT를 제안한 바 있다. 부모와 형매의 데이터를 이용하는 방법으로는 TDT와 S-TDT를 결합하여 형매의 정보를 활용하는 방법으로 C-TDT (Spielman과 Ewens, 1998)가 있다. 이 밖에도 Xu와 George (2007), Martin 등 (2000)의 방법등이 부모와 형매의 데이터를 이용하는 방법이다. 한편 Ho와 Bailey-Wilson (2000)은 X-염색체상의 마커에 대한 TDT를 제안하였으며, Horvath 등 (2000)은 S-TDT, RC-TDT를 X-염색체상의 마커에 대해 확장한 XS-TDT와 XRC-TDT를 제안하였다.

일배체형을 이용한 가족기반의 연관성 분석으로 Clayton (1999)은 일배체형을 추정하고 가능한 모든 경우에 대한 우도를 구축하는 방법을 제안하고 TRANSMIT이라는 프로그램을 개발하였는데 이는 집단의 층화에 로버스트하지 않은 것으로 알려졌다. Zhao 등 (2000)은 TDT형태의 방법을 확장하여 전달/비전달에 대한 분할표를 만들고 이의 대칭성을 검정하는 방법으로 일배체형에 대한 연관분석방법을 제안하였다. SAS/Genetics에서는 이 방법에 의한 분석을 제공한다. Horvath 등 (2004)은 가중화된 조건부 접근방법을 통해서 집단의 층화나 표현형의 분포에 민감하지 않은 검정방법을 제안하였으며 FBAT 프로그램으로 구현하였다.

2.3.2. 양적 형질인 경우 양적 형질일 때 가족기반자료의 분석방법으로서 Allison (1997)은 회귀모형에 기초한 다섯 가지의 통계량을 제시하였으며, George 등 (1999), Zhu와 Elston (2001)도 유사한 방법을 제시하였다. Fulker 등 (1999)는 부모의 정보가 없고 형매쌍에 대한 정보가 있을 때의 방법을 제시하였으며, Allison 등 (1999)도 이러한 경우에 대해 형매쌍의 수나 유전자좌의 수, 유전자간 교호작용, 공변량등을 고려할 수 있는 방법을 개발하였다. 또한 Abecasis 등 (2000)는 핵가족에 대한 양적형질 연관성 검정의 일반화방법으로서 부모의 정보여부나 형매의 수와 관계없이 사용할 수 있는 QTDT를 프로그램과 함께 제시하였으며(http://www.sph.umich.edu/cg/abecasis/QTDT), Monks와 Kaplan (2000)은 부모의 유전정보유무에 따른 세 가지 통계량을 제시하였다. 일배체형에 근거한 연관분석으로는 Horvath 등 (2004)의 방법이 양적형질인 경우도 가능하다.

표 3.1. FBAT에서의 연관분석관련 모듈과 지원 내용

Command	Option	Analysis	Trait/Marker
afreq		allele frequency (EM)	
hapfreq		haplotype frequency (EM)	
	-d	LD(D, D')	
fbat		fbat statistic allele frequency(EM) Mendelian error check	binary, continuous, censored trait
		multivariate trait test (FBAT-GEE)	multiple trait
	-t	linear combination test for multiple traits	
	-m	multi-marker test (FBAT-MM)	multi-marker
	-l	linear combination test (FBAT-LC)	
	-p	min-p test (FBAT min-p)	
	-c	fbat statistic for age-onset disease	censored trait
hbat		haplotype fbat statistic haplotype frequency (EM)	binary, continuous, censored trait
	-c	hplotype fbat statistic for age-onset disease	censored trait
	-p#	compute p-value using Monte-Carlo samples	
mode		biallelic[b], multiallelic[m], both[a]	
model		additive[a], dominance[d], recessive[r], genotype[g]	
sdt		sib-based TDT (SDT)	binary trait

3. 분석도구의 비교

3.1. 주요 분석도구별 특성

3.1.1. FBAT FBAT(family-based association test)은 미 하버드대학의 생물통계학과에서 개발된 S/W로서 핵가족이나 형매자료를 비롯한 여러 형태의 가계에 대한 데이터를 사용할 수 있으며, 이진형질 및 양적 형질, 중도절단(time-to-onset)형질에 대한 분석이 모두 가능하다 (Laird, 2009). 이진/다중대립형질(biallelic/multiallelic)마커 및 일배체형에 대한 검정과 유전모형(additive/dominant /recessive)을 고려한 검정이 가능하다. 각 마커의 위치정보가 있는 map file이 있다면 전장유전체연관(genome-wide association) 분석도 가능하다. 표 3.1에 제공되는 분석이 정리되어 있으며, 3.2절에서 각 분석별로 추가로 설명할 것이다.

3.1.2. SAS/Genetics SAS의 Genetics 모듈에서는 기초분석을 위한 PROC ALLELE, PROC HAPLOTYPE을 비롯하여, 집단기반 유전적 연관성 분석을 실행하기 위한 PROC CASECONTROL, 가족기반 유전적 연관성 분석을 실행하는 PROC FAMILY 등을 사용할 수 있다 (Czika 등, 2002; SAS Institute, 2005). Version 9.1부터는 PROC htSNP를 통해 haplotype-tagging SNPs을 확인할 수 있으며, 가계 내 근교계수(inbreeding coefficient)를 계산해 주는 PROC INBREED는 SAS/STAT과 SAS/Genetics에서 모두 실행가능하다. PROC PSMOOTH에서는 마커의 윈도우 영역에 걸친 평활 p-값과 다중 검정 보정을 위한 FDR을 제공해준다 (표 3.2 참조).

3.1.3. SAGE SAGE(Statistical Analysis for Genetic Epidemiology)는 Elston group에서 개발한 프로그램으로서 총 16개의 모듈로 구성되어 있으며 (표 3.3), 기술통계부터 자료의 질 평가, 모수적/비

표 **3.2.** SAS/Genetics에서의 분석모듈과 지원 내용

Module	Analysis
PROC ALLELE	frequency (allele, genotype, haplotype) marker informativeness indices HWE LD
PROC CASECONTROL	Chi-square test (allele, genotype, haplotype) trend test (genotype)
PROC FAMILY	Mendelian error check TDT RC-TDT sib-based TDT (S-TDT, SDT, combined SDT/S-TDT), multiallelic TDT (SDT, multiallelic combined SDT/S-TDT), X-linked TDT (X-TDT, XS-TDT, XRC-TDT)
PROC HAPLOTYPE	haplotype frequency estimation (EM, Bayesian) association test (marker-trait, haplotype-trait)
PROC htSNP	haplotype tagging SNP
PROC INBREED	inbreeding coefficient (coancestry, covariance, inbreeding)
PROC PSMOOTH	smoothing p-value (Simes' method, Fisher's method, TPM), multiple testing (Bonferroni, SIDAK, FDR)

표 **3.3.** SAGE에서의 분석모듈과 지원 내용 (*;구축중)

Classe		Module	Analysis
Preliminary analysis	Summary	PEDINFO	summary statistics
		MARKERINFO	Mendelian inconsistencies
		RELTEST	relationship testing
		FREQ	allele frequency, inbreeding coefficient
		DECIPHER	haplotype frequency
	Aggregation and Segregation	FCOR	multivariate familial correlations
		SEGREG	commingling and segregation analysis
		ASSOC	heritability estimation
		AGEON	age-onset distribution (binary trait)
	IBD Analysis	GENIBD	IBD distribution estimation (single, multi-marker)
Association analysis	Family-based	TDTEX	TDT, Generalized TDT (autosomal, continuous trait)
		ASSOC	Quantitative TDT
		DECIPHER	haplotype estimation (EM; autosomal, x-linked) LD (D') haplotype-trait association test (binary trait)
	Population-based	TBD*	population-based association
Linkage analysis	Model-Free	SIBPAL	model-free analysis (sibling pair)
		LODPAL	model-free analysis (affected sib-pairs)
		RELPAL	model-free analysis (relative pair)
	Model-based	LODLINK	model-based linkage analysis (single)
		MLOD	model-based linkage analysis (multiple)
	Design	DESPAIR	linkage design (affected relative pairs)

표 **3.4.** R의 대표적인 분석모듈과 지원 내용

Package	Analysis	Version
catmap	meta-analysis (case-control, TDT)	ver. 1.6
fbat	data quality control (missing genotype, HWE, Mendelian error) frequency (allele, genotype) family-based association test	ver. 1.6.0
fbati	gene-environment, conditional gene test (family-based)	ver. 0.6.2
gap	sample size calculations kinship calculation HWE, LD, Manhattan plot, q-q plot TDT (biallelic, multiallelic) haplotype estimation, haplotype-trait association test	ver. 1.0-20
GenABEL	genome-wide association analysis	ver. 1.4-4
genetics	frequency (allele, genotype, haplotype) marker informativeness indices HWE, LD, LD plot	ver. 1.3.4
GeneticsDesign	study planning tools (power, sample size)	ver. 1.13-0
hapassoc	haplotype-based association (GLM)	ver. 1.2-2
haplo.ccs	haplotype relative risk (population-based)	ver. 1.3
haplo.stats	haplotype frequency (EM) haplotype-based association (GLM) haplotype association design (sample size, power) sequential haplotype scan association (population-based)	ver. 1.4.3
pbatR	PBAT software interface	ver. 2.1.5
powerpkg	power analysis (affected sib pair, TDT)	ver. 1.2
SNPassoc	SNPs-based whole genome association analysis descriptive statistics (missing data patterns) HWE, LD association test (haplotype-trait) GLM (binary, quantitative traits)	ver. 1.6-0
tdthap	haplotype TDT	ver. 1.1-2

모수적 연쇄 분석, 연관분석 등을 수행할 수 있다. 대화상자와 내림메뉴 방식으로 구성되었고, 모든 분석에서 디폴트 값이 정의되어 있다. 분석용 데이터의 가계 자료 형식과 변수명에 제한이 없어 입력 파일 구조에 있어서 유연성이 매우 높다 (Elston, 2008). 여기서는 사전분석을 위한 PEDINFO, MARKERINFO, FREQ 모듈과 연관성 분석을 위한 TDTEX, ASSOC, DECIPHER 모듈 위주로 설명할 것이다.

3.1.4. R R은 무료설치가 가능하고 소스도 공개되므로 사용자가 프로그램자체를 수정 가능하다는 특징을 지닌다. 짧은 갱신 주기를 지니고 있어 사용자의 요구사항 및 버그의 신속한 해결이 가능하다. 유전통계와 관련해서도 여러 패키지들이 구현되었다(http://cran.r-project.org/web/views/ Genetics.html). 집단 유전학 분석을 위한 대표적인 패키지로 genetics, gab 등이 있으며, 연관성 분석과 관련된 패키지로 gap, tdthap, powerpkg, catmap 등이 소개되고 있다. 이 외에도 가족기반 연관성 분석을 위한 fbat, fbati 패키지 등이 개발되었다 (표 3.4). R은 뛰어난 그래픽스 기능을 제공하며, JAVA나

표 3.5. 각 S/W의 사용환경비교

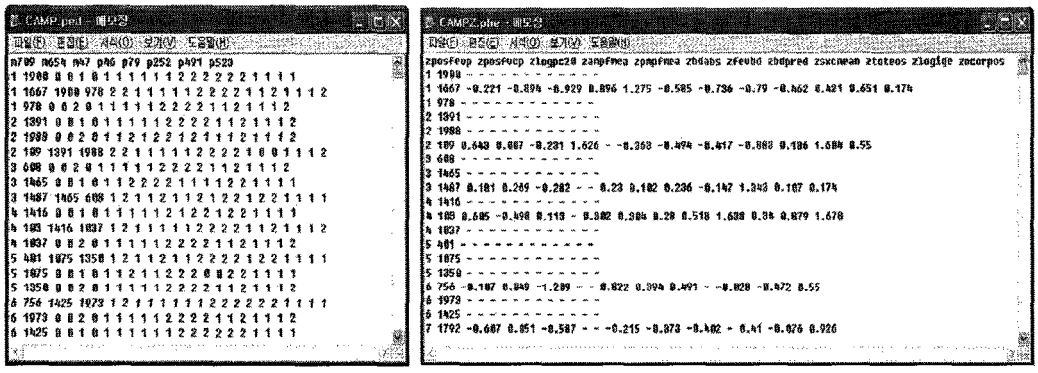
구분		FBAT	SAS/Genetics	SAGE	R
User-interface	CLI	O	O	O	O
	GUI	X	X	O	일부가능
OS	Windows	O	O	O	O
	Mac	O	O	O	O
	Unix	O	O	O	O
	Linux	O	O	O	O
유/무료		무료	유료	무료	무료
최신 version		Ver. 2.0.3	Ver. 9.2	Ver. 6.0.1	Ver. 2.9.2
웹사이트		http://www.biostat.harvard.edu/ fbat/default.html	http://www.sas.com	http://darwin.cwru.edu/sage/	http://www./r-project.org

C++ 등의 언어와 인터페이스가 가능하다. 그러나 원하는 분석을 위한 패키지를 찾기가 쉽지 않을 뿐 아니라, 비전문가에게는 프로그램 작성이 어렵다는 단점을 지니고 있다. 본 연구에서는 사전 분석 및 연관성 분석을 위한 대표적인 패키지인 genetics, gap, tdthap, hapassoc, haplo.ccs, haplo.stats, fbat 등의 패키지 위주로 설명하였다.

3.1.5. 기타 이 밖에 많이 사용되는 S/W로서 PHASE, Merlin, HAPSTAT, SNPalyze 등이 있다. PHASE는 베이지안 통계방법을 사용하여 친척관계가 없는 개체의 유전자형 자료로부터 일배체형을 재구성하고, 재조합율을 추정하며 결측된 유전자형 자료를 추정, 대체할 수 있다 (Stephens 등, 2001). 역시 베이지안 통계방법을 사용하는 fastPHASE는 PHASE보다 더 큰 자료를 다룰 수 있고 보다 빨리 결과를 출력해 주지만, 재조합율을 추정해 주지는 않는다 (Scheet와 Stephens, 2006). 두 프로그램 모두 비상업용 사용자에 한하여 무료로 다운로드 받을 수 있다(http://stephenslab.uchicago.edu/software.html). 또한 Merlin(multipoint engine for rapid likelihood inference)은 가계자료로부터 일배체형을 추정해 주며, 유전자형 오류를 점검하고 많은 마커를 가진 자료에서 모수적/비모수적 연쇄분석을 실행해 주는 프로그램으로서 양적 형질에 대한 연관성 분석, IBD 및 혈연계수(kinship coefficient) 추정 등을 수행할 수 있다 (Abecasis 등, 2002). 다양한 플랫폼에서 사용가능하고 웹사이트에서 무료 다운로드 받을 수 있다(http://www.sph.umich.edu/csg/abecasis/Merlin/index.html). 한편 HAPSTAT은 SNP 및 일배체형에 대한 연관성 분석을 위한 프로그램으로, 유전자-환경 상호작용에 대한 검정이 가능하며 윈도우와 리눅스에서 구동되고 등록을 하면 프로그램을 무료로 제공받을 수 있다(http://www.bios.unc.edu/~lin/hapstat/). SNPalyze는 마커의 다형성 정보와 HWE, LD 분석 및 EM 알고리즘을 이용한 일배체형 추정 및 사례-대조연구에서의 유전연관분석을 실행해주지만, 가족기반자료에 대한 연관분석기능은 없다. 유료프로그램이며, 웹사이트에서 데모버젼을 다운로드받을 수 있다(http://www.dynacom.co.jp/e/products/package/snpalyze/index.html.)

3.2. 분석도구별 기능 비교

3.2.1. 구동 환경 각 프로그램에 대한 사용자 인터페이스 방식과 구현 가능한 플랫폼, 지원 방식(유/무료), 최신 버전 및 홈페이지 정보는 표 3.5와 같다. FBAT과 SAS/Genetics는 CLI(command line interface) 방식인 반면, SAGE는 CLI 또는 GUI(graphical user interface) 방식으로 구현되며, R은 대부분 CLI이나 일부 패키지(fbati, pbatR)에서 CLI 또는 GUI 방식으로 구현된다. 모든 프로그



(a) 가계 파일(*.ped) (b) 표현형 파일(*.phe)

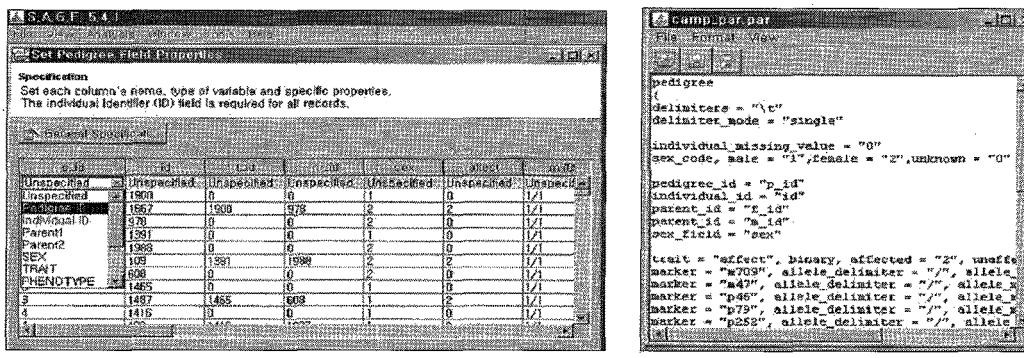
그림 **3.1.** FBAT에서 입력파일의 예

램이 다양한 플랫폼(Windows/Mac/UNIX/Linux)에서 구동가능하다. FBAT, SAGE, R은 무료로 제공되며 홈페이지에서 다운로드 받을 수 있다.

3.2.2. 데이터의 입력과 출력 FBAT의 기본 입력파일은 가계자료 파일 *.ped와 표현형자료 파일 *.phe이다. *.ped의 첫 행에는 마커의 변수명을 나열하고, 둘째 행부터 pedigree id, individual id, father id, mother id, sex, affection status, marker1_allele1, marker1_allele2,...의 순서로 한 마커당 두 열씩 연속적으로 입력되어야 한다. 자료의 부모정보가 없을 때 father id와 mother id는 0을 입력한다. sex는 남자를 1, 여자를 2로 하고, affection status는 질병이 없을 때 1, 있을 때 2, 알 수 없을 때 0으로 입력한다. 각 마커의 대립형질은 정수로 입력하되, 결측일 때는 0을 입력한다. 전장유전체연관분석을 하기 위해서 각 마커의 위치정보가 들어 있는 map file(*.map)을 넣을 수도 있는데, map file이 사용되면 가계자료 파일의 첫 행에 마커 이름이 입력되어서는 안 된다. 형질은 가계 자료 파일의 affection status이고, 양적 형질이나 중도절단 형질 등은 *.phe에 입력해야 하며 여기서도 첫 행에 형질들의 이름만을 나열하고 다른 변수명은 쓰지 않는다. 두 번째 행 이후에는 순서대로 pedigree id, individual id, trait1, trait2,...의 정보를 가계 자료 파일과 ID의 정보가 일치하도록 입력되어야 한다. 각 자료는 공백으로 구분한다 (그림 3.1 참조).

SAS/Genetics에서 연관성분석을 위한 입력양식의 경우 유전자형 자료는 한 마커 당 두 개의 변수(대립형질)로 구성되어 있으며, 각 대립형질의 자료는 각 대립형질의 자료를 숫자 또는 문자로 표현하는데 형식이 동일해야 한다. 대립형질을 나타내는 두 개의 변수 대신에 유전자형의 형태로도 입력가능하나 "/"와 같은 구분자가 필요하다. PROC ALLELE과 PROC HAPLOTYPE은 대립형질 또는 유전자형의 자료만 있어도 되지만, 집단기반 연관성 분석을 할 때는 질환의 유무를 나타내는 형질 변수가 필요하다. 가족기반 연관성 분석을 위해서는 individual id, father id, mother id가 추가로 필요한데 individual id가 가계간에 같다면 pedigree id가 있어야 하며 모든 개체에서 id 변수들의 값이 있어야 한다. 결측된 경우는 "."으로 표시하고 두 대립형질 중 하나라도 없으면 해당하는 마커의 유전자형이 결측된 것으로 간주한다.

SAGE의 기본 입력파일은 가계 파일(*.ped)로 문서파일 또는 엑셀파일에 저장된 자료를 불러올 수 있다. 가계 파일에는 개인의 기본 정보를 나타내는 pedigree id, individual id, parent 1, parent 2, sex와 표현형질, 마커, 공변량 등 모든 자료가 포함된다. 가계 파일의 변수명은 사용자가 임의로 정의할 수



(a) 가계 파일(*.ped) (b) 파라미터 파일(*.par)

그림 **3.2.** SAGE에서 입력파일의 형태

표 **3.6.** 각 S/W의 입출력 형태 비교

			FBAT	SAS/Genetics	SAGE	R
Input	Basic		*.ped, *.phe, (*.map)	*.sas7dbt	*.ped, *.par	*.* (population) *.ped, *.phe (family)
	External	Text file	O	O	O	O
		Excel file	X	O	O	O
Output	Basic		*.log	*.lst	*.out, *.inf	*.*
	External	Text file	X	O	X	O
		Excel file	X	O	X	O
	Graph		X	X	X	O

있으며 가계 파일을 불러올 때 각 변수에 해당되는 변수의 분류(pedigree id, individual id, parent 1, parent 2, sex, trait, phenotype, covariate, marker, allele, trait marker, text로 구성)를 선택하게 되므로 FBAT 프로그램과는 달리 순서에 제한을 받지 않는다. 또한 각 변수에 해당되는 변수의 분류를 선택하는 과정에서 입력된 자료의 특성, 즉 대립형질을 구분하는 문자, 결측 자료를 나타내는 값 등에 대해 입력된 형태를 사용자가 직접 정의할 수 있어 가계자료 형태에 유연성이 매우 높은 특징을 지닌다. SAGE의 또 다른 기본 입력파일은 파라미터 파일(*.par)로, 가계 파일의 각 변수에 해당되는 변수의 분류 및 자료의 입력된 형태를 정의할 때 자동으로 생성되며 사용자가 정의한 가계 파일에 대한 정보를 포함하고 있다 (그림 3.2). 파라미터 파일의 또 다른 기능은 사용자가 변수의 변환 등을 위해 함수를 작성하고자 할 때 직접 프로그래밍 할 수 있다는 것이다. SAGE의 분석 결과는 기본적으로 출력파일(*.out)로 저장되며(모듈에 따라 *.out 외 추가적인 출력파일 생성됨), 분석에 대한 경고 및 프로그램 오류 등은 정보 파일(*.inf)에 저장된다.

R 프로그램의 입출력 파일은 엑셀 및 문서파일 형태 모두 가능하며, 다양한 형태로 저장된 SNP 마커 자료(예: 대립형질 단위: 1 1, 유전자형 단위: 1/1, 대립형질 1의 개수: 2)에 대해 함수 및 인수를 이용하여 변환이 가능하다. 가족 기반 연관성 분석을 위한 입력자료 형태는 FBAT 프로그램과 동일하다. 표 3.6은 각 S/W의 입출력형태를 정리한 것이다.

3.2.3. 사전분석 FBAT은 각 마커의 대립형질 빈도와 마커들의 일배체형 빈도를 추정해 준다. 또한 두 마커 간 LD를 추정해 주며, 멘델의 법칙을 따르는지 점검할 수 있다. SAS/Genetics는 PROC AL-

표 3.7. 각 S/W에서 지원하는 사전분석방법비교

Preliminary Analysis			FBAT	SAS/Genetics	SAGE	R
Mendelian error			O	O	O	O
Frequency estimation	allele		O	O	O	O
	genotype		O	O	X	O
	haplotype	proportion	O	O	O	O
		individual assign	X	O	O	O
Informativeness index	PIC		X	O	X	O
	Het		X	O	X	O
HWE			X	O	X	O
LD	D		O	O	X	O
	D'		O	O	O	O
	r		X	O	X	O
	LD plot		X	X	X	O

LELE에서 각 마커의 정보력지수를 계산해주고, HWE에 대해 카이제곱 적합도검정과 Guo와 Thompson (1992) 검정을 출력한다. 또한 두 마커 간 연관성에 대한 척도인 LD를 추정하지만 LD plot과 같은 그림으로 제공하지는 않는다. PROC HAPLOTYPE에서는 각 개인의 유전자형으로 구성할 수 있는 가능한 일배체형쌍을 개별로 제시하고 그에 대한 확률을 제공한다. SAGE의 PEDINFO모듈은 가계 자료에서 가족, 형매, 가계크기에 대한 기술통계량 및 친척 관계의 유형에 따른 수를 계산한다. MARKERINFO모듈에서는 멘델 법칙에 따르지 않는 자료에 대한 정보를 제공하고, FREQ모듈에서는 가계자료를 이용하여 대립형질 빈도 및 대립형질 빈도의 최대우도추정량를 계산한다. 유전자형 빈도는 현재 제공되지 않으며, 일배체형 빈도 추정은 DECIPHER모듈에서는 제공한다. R에는 사전분석을 지원하는 패키지들이 여러 개 존재하는데, genetics 패키지에서 유전자형 및 일배체형 자료의 형태를 정의하기 위한 함수들이 제공되며, 대립유전자 빈도 추정, HWE 및 LD에 대한 추정 및 검정 등이 가능하다. haplo.stats 패키지에서는 일배체형 확률의 최대우도추정량 및 EM 알고리즘을 이용한 최대우도추정량 계산이 가능하다. 또한 SNP 마커 간 LD 정보의 시각화 뿐 아니라 사전 분석 결과에 대한 그래프화가 가능하다. 표 3.7에 각 S/W별로 지원되는 사전분석방법을 정리하였다.

3.2.4. 집단기반 연관성분석 FBAT 패키지는 집단기반 연관성분석을 지원하지 않고 있다. SAS/Genetics는 이진 및 다중 대립형질 마커간 연관성 분석을 할 수 있으며, CMH 검정을 통한 층화분석이나 오즈비를 구할 수 있다. 또한 일배체형의 빈도를 추정해서 그 비율이 사례-대조군간에 다른지와 각 일배체형과 그 일배체형을 제외한 나머지 일배체형의 비율이 두 군간 다른지 검정할 수 있다. SAGE 프로그램의 집단기반 분석을 위한 모듈 TBD는 개발 중으로 아직 제공되지 않는다. R의 gap 패키지는 집단기반 연관성 연구에서 요구되는 표본수 계산이 가능하며 SNP 및 일배체형과 표현형질간의 연관성 분석을 위한 스코어 검정 (Schaid 등, 2002), 일배체형의 추세 회귀분석 (Zaykin 등, 2002; Xie와 Stram, 2005) 등을 지원한다. haplo.stats 패키지는 일배체형을 설명변수로 하는 이진형질 및 양적형질의 GLM 분석이 가능하다 (Lake 등, 2003). GLM 분석시 유전모형을 고려할 수 있으며 유전자와 공변량간의 상호작용에 대한 분석도 제공한다. haplo.ccs 패키지는 사례-대조 연구에서 가중로지스틱 회귀를 이용하여 일배체형 및 공변량의 상대위험도를 추정한다 (French 등, 2006). SNPassoc은 전장유전체연관분석을 위한 패키지로 이진 또는 양적형질 및 5가지 유전모형(codominant/dominant/recessive/overdominant/log-additive)에 따라 GLM에 기초한 연관성분석을 실시하며, 일배체형에 대한 연관성 분석도 가능하다 (Gonzalez 등, 2007). 표 3.8은 집단기반 분석

표 **3.8.** 각 S/W에서 지원하는 집단 기반 연관분석방법비교

Population-based Analysis		FBAT	SAS/Genetics	SAGE	R
Binary Trait	Chi-square test	X	O	X	O
	Exact test	X	X	X	O
	Trend test	X	O	X	O
	Logistic regression	X	X	X	O
Quantitative Trait	Regression	X	X	X	O
	ANOVA	X	X	X	O
	GLM with interaction	X	X	X	O

에 대해 제공되는 분석을 정리한 것이다.

3.2.5. 가족기반 연관성분석 FBAT 패키지는 명령어 fbat과 함께 사용되는 옵션을 사용하여 이진, 연속 및 중도절단형질에 대한 연관검정을 모두 수행할 수 있다. TDT외에 형매자료에 대한 분석으로 S-TDT를 제공하며, 다중 마커검정으로서 Rakovski 등 (2007)의 FBAT-MM(-m옵션)과 Lange 등 (2003)의 FBAT-LC(-l옵션)이 실행되며, -t 옵션이나 FBAT-GEE를 사용하여 다변량분석도 가능하다. 또한 이진형질뿐 아니라 양적형질에 대해서도 분석가능하며, 데이터를 불러올 때(load command) -x 옵션을 함께 써주면 X-염색체상의 마커도 분석할 수 있다. 명령어 hbat으로는 이진 및 연속, 중도절단 형질에 대해 일배체형과의 연관검정이 가능하다. FBAT에서 공변량을 보정하고자 할 때에는 프로그램의 외부, 즉 SAS나 기타 소프트웨어로부터 공변량과 마커를 독립변수로 하고 종속변수를 이진 형질 또는 양적 형질로 하는 로지스틱 회귀모형 또는 선형회귀모형을 적합한 뒤 그 잔차를 표현형 자료 파일에 새로운 형질로 입력해서 fbat 명령어를 구동해야 한다.

SAS/Genetics의 경우 가족기반 자료에서 마커와 이진형질간 연관성을 TDT, S-TDT, SDT, RC-TDT로 검정할 수 있고, 가족의 형태에 따라 앞서의 검정방법들을 각각 적용한 다음 검정통계량을 결합하여 검정하는 것도 가능하다(combined test). 또한 XLVAR문을 이용하면 X-linked 마커에 대해서도 위 검정들을 모두 실행할 수 있다. 일배체형의 경우 EM과 베이즈절차를 이용한 추정과 검정을 수행한다. 양적형질에 대해서는 분석결과를 제공하지는 않으며, 다만 PROC FAMILY에서 outq=dataset옵션을 이용하여 각 마커의 대립형질에 대한 전이점수(transmission score)를 받은 다음, SAS/STAT에서 계산절차를 연구자가 직접 프로그래밍하는 방식으로만 가능하다.

SAGE의 TDTEX는 TDT 분석을 수행하는 모듈로 상염색체의 마커에서 이진형질에 대한 분석으로 제한된다. 가족 당 질환을 가진 자손이 1명인 경우에 대한 고전적 TDT분석 뿐 아니라 질환을 보인 자손이 여러 명인 경우, 2명의 자손 중 1명이 질환자인 경우, 자손 중 여러 명이 질환자인 경우 등의 확장된 가계 자료 및 다중대립형질 마커 자료에 대한 일반화 TDT (Curtis와 Sham, 1995; Rice 등, 1995) 분석도 지원된다. TDTEX 모듈을 이용한 분석 시 옵션을 이용하여 분석에 사용될 가계와 질환을 보인 자손의 수 및 질환을 보인 형매의 수를 옵션을 통해 지정할 수 있다. ASSOC 모듈은 가족 상관이 존재하는 확장된 가계 자료에서 양적 형질과 한 개 이상의 공변량 간의 연관성 분석을 실시하며 (Elston 등, 1992; George 등, 1999), 연관성 분석을 위해 가족 상관 구조 및 회귀모형을 가정한 후 모수가 추정된다. 공변량으로 연속형 자료로 변환된 마커 자료가 포함될 수 있으며, 우도비 검정을 통해 공변량에 대한 유의성을 평가한다. DECIPHER에서는 이진형질에 대해서 일배체형과의 연관성 검정을 제공한다.

R 프로그램에서 가족기반 연관성분석을 위한 패키지는 아직 제한적이다. 대표적인 패키지로는 fbat, fbati, gap, pbatR, tdthap, powerpkg가 있다. fbat 패키지는 FBAT 프로그램에서 수행되는 분석 중 일부를 지원하고 있으며, 연관성 분석을 위한 가계 자료에서 가족단위는 핵가족이어야 하며 자손의 유전

표 3.9. 각 S/W에서 지원하는 가족 기반 연관분석방법비교

Family-based Analysis			FBAT	SAS/Genetics	SAGE	R
Binary trait	SNP	Basic TDT	O	O	O	O
		Sib-based TDT	O	O	O	O
		Combined TDT	O	O	O	O
		X-linked TDT	O	O	X	O
		Multiallelic TDT	O	O	O	X
	haplotype	association test	O	O	O	O
Quantitative trait	SNP	Basic TDT	O	X	O	X
		Sib-based TDT	O	X	O	X
		Combined TDT	O	X	O	X
		X-linked TDT	O	X	X	X
		Multiallelic TDT	O	X	O	X
	haplotype	association test	O	X	X	X

자 및 표현형 자료에서 결측이 없어야 하고 2개의 대립형질을 가진 마커에 한해서 유전모형에 따른 fbat 통계량을 계산할 수 있다. fbati 패키지는 이진대립형질 마커에 대해 유전자-환경 상호작용 검정을 수행한다 (Lunetta 등, 2000). 또한 gap 패키지는 가족기반 연구(TDT, ASP, ASP-TDT)의 설계에서 요구되는 표본 크기를 계산해 주고 (Risch와 Merikangas, 1996), 대립형질이 2개 이상인 마커에 대한 TDT 및 QTL 등을 지원해 준다 (Sham, 1998). pbatR은 R 프로그램 내에서 PBAT 프로그램과의 인터페이스를 위한 패키지로, PBAT 프로그램의 출력결과를 자동적으로 읽을 수 있으며 그림을 출력할 수 있다. tdthap은 일배체형에 대한 TDT 분석, powerpkg는 TDT 및 ASP에서 연관검정을 위한 Power를 추정한다. 표 3.9는 각 S/W별로 가족기반연관분석의 지원여부를 정리한 것이다.

4. 토의

유전적 연관성 분석을 위한 4개 프로그램을 비교하자면, FBAT의 경우 다양한 형태의 가족 자료에 대해 이진/양적/중도절단된 형질과 마커와의 연관성을 검정할 수 있다. 가족기반자료에 대해서는 가장 많이 사용되는 프로그램이라고 할 수 있으며 비교적 사용하기 쉽다. SAS/Genetics는 이진 형질에 대해 사전분석이나 집단기반 분석, 가족기반 분석을 쉽게 수행할 수 있으며, 출력된 자료를 사용하여 SAS/STAT에서 지원되는 다양한 분석을 추가로 실행할 수 있는 강점이 있다. 데이터의 입출력이나 분석결과의 보고양식이 체계적이고, 기존의 프로그램형식을 따르고 있으므로 SAS에 익숙한 통계전공자들에게는 편리한 프로그램이다. 그러나 유료이며, 분석방법의 업데이트 속도가 늦어 아직 충분한 분석방법을 확보하고 있지 못하다는 단점이 있다. SAGE는 입출력이나 분석방법 등을 메뉴방식으로 선택하여 일괄적으로 수행할 수 있으며 분석체계가 잘 잡혀있다. 가계 자료에 대한 상세한 기술통계량을 제공하여 사용자가 가계 정보를 한 눈에 볼 수 있다. 그러나 연쇄분석과 달리 연관성 분석을 위한 모듈은 아직 개발 단계에 있는 것이 많아 사용이 제한적이다. R은 분석 과정에서 함수 사용에 따른 인수들을 사용자가 자유롭게 정의할 수 있어 분석의 유연성이 높으며, 다양한 설계에 따른 표본 크기 추정 및 검정력 분석도 개발되어 있다. R의 장점은 무엇보다 빠른 업데이트가 가능하여 최신 방법을 사용할 수 있다는 점이다. R의 경우는 성격상 모든 프로그램을 망라하는 것이 불가능하다. 여기서는 CRAN의 목록에 있는 프로그램과 추가로 많이 사용되는 프로그램위주로 정리하였다.

그림 4.1은 본문에서 각 분석 내용 및 단계별로 가능한 프로그램목록을 정리한 것이다. 굵은 글씨로 표기한 것이 추천 프로그램이다. 이와 같이 각 S/W는 분석의 일부분을 제공하며 유전적 연관분석을 온전

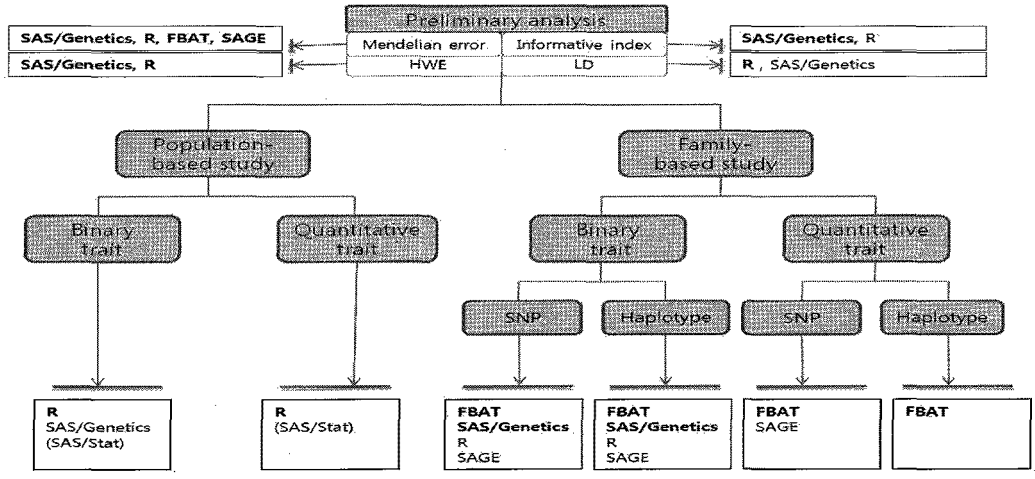


그림 **4.1.** 분석 내용 및 단계별 사용가능한 프로그램

히 통합하고 있지 못하다. 따라서 본고에서 제시된 분석표에 근거하여 각 연구자의 목적에 따라 적절한 프로그램을 선택하는 것이 필요할 것이다.

참고문헌

Abecasis, G. R., Cardon, L. R. and Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families, *American Journal of Human Genetics*, **66**, 279–292.

Abecasis, G. R., Cherny, S. S., Cookson, W. O. and Cardon, L. R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees, *Nature Genetics*, **30**, 97–101.

Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits, *American Journal of Human Genetics*, **60**, 676–690.

Allison, D. B., Hero, M., Kaplan, N. and Martin, E. R. (1999). Sibling-based test of linkage and association for quantitative trials, *American Journal of Human Genetics*, **64**, 1754–1764.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies, *Nature Reviews Genetics*, **7**, 781–91.

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps, *Bioinformatics*, **21**, 263–265.

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

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

Curtis, D. (1997). Use of siblings as controls in case-control association studies, *Annals of Human Genetics*, **61**, 319–333.

Curtis, D. and Sham, P. C. (1995). An extended transmission/disequilibrium Test(TDT) for multi-allele marker loci, *Genetic Epidemiology*, **7**, 319–334.

Czika, W., Yu, X. and Wolfinger, R. D. (2002). A introduction to genetic data analysis using SAS/Genetics, SAS Institute Inc., Cary, North Carolina, USA.

Elston, R. C. (2008). Statistical analysis for genetic epidemiology(S.A.G.E.) user reference manual (Version 5.4.2), Case Western Reserve University, Cleveland, Ohio.

Elston, R. C., George, V. T. and Severtson, F. (1992). The Elston-Stewart algorithm for continuous genotypes and environmental factors, *Human Heredity*, **42**, 16–27.

Ewens, W. J., Li, M. and Spielman, R. S. (2008). A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker, *PLoS Genetics*, **4**, e1000180.

Fallin, D. and Schrock, N. J. (2000). Accuaracy of haplotype frequency estimation of biallelic loci, via the expectaation-maxmization algorithm for inphased diploid genotype data, *American Journal of Human Genetics*, **67**, 947–959.

French, B., Lumley, T., Monks, S. A., Rice, K. M., Hindorff, L. A., Reiner, A. P. and Psaty, B. M. (2006). Simple estimates of haplotype relative risks in case-control data, *Genetic Epidemiology*, **30**, 485–494.

Fulker, D. W., Cherny, S. S., Sham, P. C. and Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits, *American Journal of Human Genetics*, **64**, 259–267.

George, V. T., Tiwari, H. K., Zhu, X. and Elston, R. C. (1999). A test of transmission/disequilibrium for quantitative traits in pedigree data by multiple regression, *American Journal of Human Genetics*, **65**, 236–245.

Gonzalez, J. R., Armengol, L., Sole, X., Guino, E., Mercader, J. M., Estivill, X. and Moreno, V. (2007). SNPassoc: an R package to perform whole genome association studies, *Bioinformatics*, **23**, 654–655.

Guo, S. W. and Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles, *Biometrics*, **48**, 361–372.

Hirshhorn, J. N., Lohmueller, K., Byrne, E. and Hirshchhorn, K. (2002). A comprehensive review of genetic association studies, *Genetics in Medicine*, **4**, 45–61.

Ho, G. Y. F. and Bailey-Wilson, J. E. (2000). The transmission/disequilibrium test for linkage on the X chromosome, *American Journal of Human Genetics*, **66**, 1158–1160.

Horvath, S. and Laird, N. M. (1998). A discordant-sibship test for disequilibrium and linkage: No need for parental data, *American Journal of Human Genetics*, **63**, 1886–1897.

Horvath, S., Laird, N. M. and Knapp, M. (2000). The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers, *American Journal of Human Genetics*, **66**, 1161–1167.

Horvath, S., Xu, X., Lake, S. L., Silverman, E. K., Weiss, S. T. and Laird, N. M. (2004). Family based tests for association haplotypes with general phenotype data: Application to athma genetics, *Genetic Epidemiology*, **26**, 61–69.

Knapp, M. (1999). The transmission/disequilibrium test and parental-genotype reconstruction: The reconstruction-combined transmission/disequilibrium test, *American Journal of Human Genetics*, **64**, 861–870.

Laird, N. M. (2009). Family-based association tests and the FBAT-toolkit user's manual (updated march 2009), Harvard school of public health, Boston, MA.

Laird, N. M. and Lange, C. (2008). Family-based methods for linkage and association analysis, *Advances in genetics*, **60**, 219–252.

Lake, S., Silverman, E., Weiss, S., Laird, N. and Schaid, D. J. (2003). Estimation and tests of haplotype environment interaction when linkage phase is ambiguous, *Human Heredity*, **55**, 56–65.

Lange, C., Silverman, E. K., Xu, X., Weiss, S. T. and Laird, N. M. (2003). A multivariate family-based association test using generalized estimating equations: FBAT-GEE, *Biostatistics*, **4**, 195–206.

Lunetta, K., Faraone, S. V., Biederman, J. and Laird, N. M. (2000). Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions, *American Journal of Human Genetics*, **66**, 605–614.

Martin, E. R., Monks, S. A., Warren, L. L. and Kaplan, N. L. (2000). A test for linkage and association in general pedigrees: The pedigree disequilibrium test, *American Journal of Human Genetics*, **67**, 146–154.

Monks, S. A. and Kaplan, N. L. (2000). Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus, *American Journal of Human Genetics*, **66**, 576–592.

Monks, S. A., Kaplan, N. L. and Weir, B. S. (1998). A comparative study of sibship tests of linkage and/or association, *American Journal of Human Genetics*, **63**, 1507–1516.

Rakovski, C., Xu, X., Lazaras, R. and Laird, N. (2007). A new multimarker test for family-based association studies, *Genetic Epidemiology*, **31**, 9–17.

Rice, J. P., Neuman, R. J., Hoshaw, S. L., Daw, E. W. and Gu, C. (1995). TDT with covariates and genomic screens with mod scores: their behavior on simulated data, *Genetic Epidemiology*, **12**, 659–664.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases, *Science*, **273**,

1516–1517.

Saito, Y. A., Talley, N., Andrade, M. and Petewrsen, G. (2006). Case-control genetic association studies in gastrointestinal disease: Review and recommendations, *American Journal of Gastrointorology*, **101**, 1379–1389.

SAS Institute. (2005). SAS Genetics 9.1.3 User's Guide, SAS Institute, Inc. Cary, NC.

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, *American Journal of Human Genetics*, **70**, 425–434.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase, *American Journal of Human Genetics*, **78**, 629–644.

Sham, P. C. (1998). Statistics in Human Genetics, Arnold.

Slager, S. L. and Schaid, D. J. (2001). Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects, *American Journal of Human Genetics*, **68**, 1457–1462.

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.

Spielman, R. S. and Ewens, W. J. (1998). A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test, *American Journal of Human Genetics*, **62**, 450–458.

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.

Xie, R. and Stram, D. O. (2005). Asymptotic equivalence between two score tests for haplotype-specific risk in general linear models, *Genetic Epidemiology*, **29**, 166–170.

Xu, H. and George, V. (2007). A new transmission test for affected sib-pair families, *BMC Proceedings*, **1**(Suppl 1), S32.

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, H. (2000). Family-based association studies, *Statistical Methods in Medical Research*, **9**, 563–587.

Zhao, H., Zhang S., Merikangas, K. R., Wildenaur, D., Sun, F. and Kidd, K. K. (2000). Transmission/disequilibrium test for multiple tightly linked markers, *American Journal of Human Genetics*, **67**, 936–946.

Zhu, X. and Elston, R. C. (2001). Transmission/disequilibrium test for quantitative traits, *Genetic Epidemiology*, **20**, 57–74.

A Review of Genetic Association Analyses in Population and Family Based Data: Methods and Software

Hyo-Jung Lee[1] · Min-Ji Kim[2] · Mira Park[3]

[1]Department of Statistics, Korea University
[2]Biostatistics Team, Samsung Biomedical Research Institute
[3]Department of Preventive Medicine, Eulji University

(Received September 2009; accepted December 2009)

Abstract

Recently, there have been lots of study for disease-genetic association using SNPs and haplotypes. Statistical methods and tools for various types of data are developed by many researchers. However, there is no unified software which can handle most of major analysis, and the methods and manners to deal with data are quite different through softwares. And thus it is not easy to researcher to choose proper software. In this study, we devide analyzing procedures into three steps: preliminary analysis, population-based analysis and family-based analysis. We review the statistical methods for each step and compare the features of the FBAT, SAS/Genetics, SAGE and R as major integrating softwares for genetic study.

Keywords: Genetic association, SNP, haplotype, softwares.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund)(KRF-2007-531-C00018).

[3]Corresponding author: Associate Professor, Department of Preventive Medicine, Eulji University, Daejeon 301-832, Korea. E-mail: mira@eulji.ac.kr