원 저

韓國疫學會誌 : 第30卷 第1號
Korean Journal of Epidemiology
Vol.30, No.1, June, 2008, 49-59

가족기반 코호트 연구의 사례와 전망

설재웅[1,4], 박수경[2], 오희철[3], 지선하[1,4]

Department of Epidemiology, Johns Hopkins Bloomberg Schoolof Public Health, Baltimore, MD, USA[1],
서울대학교 의과대학 예방의학교실[2], 연세대학교 의과대학 예방의학교실[3], 연세대학교 보건대학원 국민건강증진연구소[4]

서 론

코호트 연구는 만성질환의 원인을 이해하는데 있어서 주요한 역할을 하여왔으며 그 중요성은 더욱 부각될 것이다[1]. 반면에 단면연구는 질병의 유전요인 및 환경요인을 이해하는데 있어서 일부의 성공을 거두었을 뿐 후속 연구들에서 결과가 일치하지 않는 등의 한계를 보여왔다. 연구결과 반복에 실패하는 주요 원인으로는 낮은 통계적 검정력(statistical power)과 연구대상 집단간에 이질성(heterogeneity) 등이다. 다른 가능한 이유로는 단면연구는 질병의 진행과정에 따른 원인의 다양성을 설명하지 못한다는 점이다. 예를 들어서 다른 유전자들이 동일질병의 다른 진행단계에 작용할 수 있다. 질병의 시작 (initiation)단계와 진행(progression) 단계에서 작용하는 유전자가 다를 수 있는 것이다. 환자대조군 연구의 경우는 주로 중증 환자(severe case)만을 다룬다는 단점 이외에도 환자군과 대조군간에 생활습관요인에 대한 회상에 차이가 나는 회상 바이어스(recall bias)의 문제를 추가적으로 갖는다[1]. 따라서 만성질환의 유전 및 환경요인의 연구를 위해서는 장기간의 추적기간을 가지는 코호트를 통한 연구가 필요하다. 코호트 연구는 일반적으로 독립적인 대상자(unrelated subjects)를 가진 일반인구 코호트가 많이 이용되지만 이 경우는 가족간에 공유되는 환경요인이나 유전요인에 대한 보정(adjustment)이 가능하지 않다[2]. 또한 최근 연구가 많이 이루어지고 있는 imprinting (parent-of-origin) effect에 대한 분석도 가능하지 않다. 따라서 가족간의 환경요인과 유전요인의 보정이나 유전자 탐색을 위해서는 가족 연구가 유일한 방법이다.

이 연구에서는 세계적으로 가족 코호트의 연구현황과 문제점을 정리하고, 앞으로의 전망과 방향을 모색하고자 한다. 가족기반 코호트 연구의 대표적인 사례로는 1948부터 시작된 프램험 코호트(Framingham cohort)를 기반으로 한 가족코호트이다. 프램험 연구(Framingham study)를 중심으로 하여 가족기반 코호트 연구와 가족연구의 사례와 전망을 요약하였다. 현재 이루어지고 있는 대표적인 가족기반 코호트 연구와 가족연구는 표1과 같다 (표1). 여기서 가족연구는 과거의 "질환 중심가계"연구와 가족기반 코호트 연구 (family-based cohort 또는 population-representative family)를 모두 포함하였다.

접수 : 2008년 4월 1일 채택 : 2008년 5월 23일
교신저자 : 지선하
주소 : 연세대학교 보건대학원 국민건강증진연구소
서울특별시 서대문구 신촌동 134
전화 : +82-2-2228-1523 팩스 : +82-2-365-5118
E-mail : jsunha@yuhs.ac
이 과제는 질병관리본부 학술연구용역사업과 (2007-E00004-00) 서울시 산학연협력사업의 지원에 의해 이루어진 것임(10526).

가족기반 코호트와 가족연구의 사례

1. 프램험 연구

프램험 심장 연구(The Framingham Heart Study)는 1948년에 28세에서 62세 사이의 성인 5,209명(남성 2336명, 여성 2873명)을 대상으로 미국 메사추세스, 프램험 지역에서 시작되었다. 프램험 지역 가구의 3분의 2가 계통적 표본 추출 (systematic sample) 로 선정되었다. 연구 시작 단계에서 가족 연구로 계획된 것은 아니지만 1,644명의 배우자 쌍이 연구에 포함되었다. 이들 대상자들은 매 2년마다 심장질환에 대한 추적이 이루어졌다.

1971년에는 기존 코호트 대상자의 자녀와 배우자 5,124명을 대상으로 2세대 코호트 자료가 수집되었다.

Table 1. Examples of family-based study

No.	Country	Starting Year	Study Name	Study Phenotype	No. of Subjects	Study design
1	USA	1948	Framingham SHARe	8 phenotypes group	15,876	Community-based, longitudinal, family based cohort
2	EU 8 countries, and USA	1999	International Multi-Center ADHD Genetics Project	ADHD	2,835	Parent-offspring trios
3	Australia	1991	The Victorian Family Heart Study (VFHS)	CVD	2,911 (767 families)	Family based cohort
4	USA	1995	The Amish Family Diabetes Study	Diabetes	617	Family Study
5	USA	1991	The NIMH Alzheimer's Disease Study	Alzheimer's Disease	1439 (437 families)	Family based cohort
6	USA, Canada, and Australia	1995	The Breast Cancer Family Registry	유방암	11,950 families	Cancer Family Registry

1:Cupples et al.[3], 2:Kuntsi et al.[4], 3:Ellis et al.[5], 4:Pollin et al.[6], 5:Bertram et al.[7], 6:John et al.[8]

이들 자녀 코호트는 매 4년마다 추적되었다. 1980년대 말부터는 DNA를 연구대상자로부터 수집하였다. DNA는 1948년 당시 대상자들이 대부분 사망하였으므로 주로 자녀(2세대) 코호트에서 얻어졌다. 1990년 중후반에는 DNA 자료가 330 가족에서 얻어졌다. 이 330 가족자료에서는 3,041 부모-자녀쌍, 2,796 자녀쌍, 1,595 사촌쌍 등이 포함되었다. 2000년 초에는 이 330가계도로부터 1,399명을 대상으로 Affymetrix 100K GeneChip을 이용한 유전자형 분석(genotyping)이 이루어졌다. 유전자형(Genotype) 자료 처리(cleaning) 이후에 1,345명의 코호트 대상자 자료 (278명의 기존 1948년 코호트, 1,087명의 자녀 코호트)가 남았다.

2002년부터는 1948년 기존코호트의 손자, 손녀를 대상으로 3세대 코호트 연구를 시작하여서2005년까지 4,095명의 대상자가 참여하였다. 프램험 연구의 표현형(phenotype)은 크게 8군으로 분리되어서 혈압, 동맥경직성(arterial stiffness), 운동능력시험, 대사 관련 변수, 폐기능 및 수면 변수, 감염 관련 바이오 마커, 잠재성 동맥경화증(subclinical atherosclerosis), 신장 및 내분비 기능 변수, 장수 및 노화 관련 변수, 심장질환, 암 등이 조사되었다.

이 프램험 연구를 통해서 현재까지 무려 1,731편의 논문이 국제저널에 출판되었다. 대다수는 기존 코호트 자료를 이용한 심혈관질환의 생활습관 원인을 연구한 것이고, 100여편의 linkage 분석 논문을 포함하여 많은 심혈관질환 유전역학 연구가 이루어져왔다. Linkage 분석 이외에도 후보 유전자와 심혈관질환의 관련성을 연구하는 후보유전자 연구방법(candidate gene approach)이 이루어져 왔다. 그 예로 2003년에는 1739명의 자녀 코호트에서 ESR1 c.454-397T>C polymorphism과 심혈관 질환의 관련성 연구를 하였다. 이 연구에서 CC 유전자형(genotype)을 가진 사람들이 CT나 TT 유전자형(genotype)을 가진 사람들에 비하여서 전체 심혈관질환에 걸릴 위험이 2배 높았고(p value=0.004), 심근경색에 걸릴 위험은 3배가 높았다(p value=0.001)[9].

프램험 연구에서 최근에는 Whole-genome wide association 방법을 이용한 심혈관 질환 유전역학 연구가 활발히 이루어지고 있다. 2006년에는 프램험 자녀 코호트 대상자 694명에 대하여116,204 SNPs에 대한 유전자형(genotype) 분석을 하여 비만 관련 유전자에 대한 연구를 발표하였다. 이 연구에서는 비만을 연속형 변수(quantitative traits)로 두고 PBAT 소프트웨어의 "conditional mean model"을 이용하여 다중비교(multiple testing)의 문제를 해결하기 위한 방법으로 2단계 접근(two-stage testing)을 하였다. 즉, 첫 번째 screening 단계에서는 실제 표현형 분석의 대상인 2세의 유전형은 사용하지 않고, 그 부모의 유전자형 (genotype)만을 사용하여, 모든 가능한 2세들의 유전형들과 관련성을 가질 수 있는 SNP들의 통계적 검정력(power)을 계산하였다. 2단계에서는 1단계에서 검정

력이 높은 SNP들만을 가지고, 2세대의 유전형을 그대로 사용하여 비만이라는 표현형과 관련성을 보이는 SNP을 분석하였다. 즉, 1단계에서는 실제 분석에 사용되는 2세들의 유전형을 전혀 사용하지 않고 screening을 하였기 때문에, 다중비교에서 자유로우며, 예를 들어, 전체 50만개의 SNP 중에서 100개의 높은 power를 가진 SNP만으로 2단계 분석이 이루어졌다면, 다중비교의 보정은 50만개가 아닌 100개에 대해서만 수행하면 된다는 것이다. (즉, p-value가 5x10-4 수준) 이처럼 전체 연구의 power의 손실 없이 다중비교의 단점인 over-correction의 문제를 해결할 수 있는 것은 unrelated population으로 이루어지는 환자-대조군 연구에서는 원천적으로 불가능한 가족 연구의 장점이다. 이 연구의 FBAT 분석에서는 2번 염색체의 SNP rs7566605만이 P-value 0.0026로 유의하였다. 연구팀에서는 이 SNP에 대하여 다른 인구에서 얻어진 연구자료를 통해 결과를 검증하였다. 예를 들어서 독일의 KORA 코호트의 3,996명에 대한 분석에서는 이 SNP에 대한 Linear regression 분석에서p-value 0.008을 얻었다. 또한 미국 Maywood 지역의 African-American 가족 연구에서도 866명의 대상자에 대한 PBAT 분석에서 동일한 SNP에서 p-value 0.009를 얻어서 결과를 확인하였다[10].

또한 2007년에는 100K GeneChip 자료를 이용한 논문 17편이 동시에 출판되었다[3]. 이 연구는 프램험 가족기반 코호트를 이용하여 17개의 표현형(phenotype)에 대해서 17편의 논문을 동시에 만든 것으로서 코호트 연구의 장점을 잘 보여주는 예이다. 이 연구에서는 linkage분석, FBAT 분석과generalized estimating equation (GEE) 분석을 하였다. 주요 연구결과는 표2와 같다. 연구팀은 linkage 분석과association 분석 결과가 일치하는 결과를 보고하였다. 그 예로 monocyte chemoattractant protein-1에 대하여 염색체 1번에서 LOD score=4.96의 유의한 결과를 얻었는데 그와 근접한 OR10J1과 OR10J3 유전자rs4128725와 rs2494250 SNP에서 유의한 결과를 얻었다. 또한 factor VII과 유의한 관련성을 보인 13번 염색체의 rs561241 SNP의 경우는 이미 이전의 다른 연구에서 관련성이 보고된 SNP이다. 또한 프램험 연구팀은 2007년에는 9,300명의 연구대상자를 대상으로 550,000 SNPs 분석을 하여서 추가 Whole-genome wide association 연구를 진행중이다.

2. International Multi-Center ADHD Genetics (IMAGE) Project

IMAGE project은 유럽의 8개 국가의 12개 기관에서 주의력 결핍 과잉 행동장애(Attention Deficit Disorder with Hyperactivity ADHD) 질환을 대상으로 958 부모 자녀 트리오 (parent-child trios)의 자료를 수집한 다국적 공동 연구이다. 유럽의 8개 국가 이외에 미국의 하버드 보건대학원팀도 통계분석에 참여하고 있다. 연구팀은 추가적으로 1,400 가족의 자료까지 모을 계획을 가지고 있다[4]. ADHD는 어린이의 3-10%, 성인의 2-4%의 유 병률 을 갖는 흔한 신경발달(neurodevelopmental) 질환이다. 6-17세의 ADHD 환자와 그 형제, 그리고 부모들이 대상에 포함되었다. 연구팀은 2006년에 51개 유전자에 위치한 1,038개 single-nucl-

Table 2. Results of Framingham 100K GeneChip

Phenotype working group	Trait	SNP rs ID	Chr	GEE P-value	FBAT P-value	IN/NEAR gene
Select biomarkers	Monocyte chemoattractant protein-1	Rs2494250	1	1.0*10-14	3.5*10-8	FCERIA, ORIOJ3
	Monocyte chemoattractant protein-1	Rs4128725	1	3.7*10-12	3.3*10-8	ORIOJ1
Kidney/Endocrine	Cystatin C	Rs1158167	20	8.5*10-9	0.006	CST9L/CST9
Diabetes	fasting plasma glucose	Rs2722425	8	2.0*10-8	0.005	ZMAT4
Neruology	Total Cerebral Brain Volume (ATCBV)	Rs1970546	20	4.0*10-8	0.005	CDH4
Hemostatic factors	Factor VII	Rs561241	13	4.5*10-16	3.4*10-4	F7

Source: Cupples et al.[3]

Table 3. Haplotype analysis using 5-SNP sliding window method and analyses using UNPHASED

Gene	Marker Window	P-value	Transmitted	Non-Transmitted	OR	Haplotype-specific P-value
NET1	16-17-18-19-20	0.005	119	95	1.25	0.101
TPH2	36-37-38-39-40	0.007	206	151	1.36	0.004
PER2	3-4-5-6-7	0.016	188	160	1.18	0.133
ADRB2	4-5-6-7-8	0.024	137	98	1.40	0.011
HTR1E	9-10-11-12-13	0.031	15	8	1.88	0.144
MAOA	12-13-14-15-16	0.033	167	133	1.26	0.050
CHRNA4	11-12-13-14-15	0.046	14	3	7.12	0.008

Source: Brookes et al.[11]

eotide polymorphisms (SNP)에 대한 결과를 보고하였다[11]. 표 3은 haplotype 분석의 결과이다. 다중비교(Multiple testing)로 인한 type 1 error를 줄이기 위해서 통계적으로 유의한 SNP을 가진 유전자에 대해서만 haplotype 분석을 하였다. 분석은 WHAP 프로그램을 이용하여 5개 SNP을 순차적으로 분석해나가는 sliding window 방법을 이용하여 분석하였다. WHAP 프로그램은 하버드대학 웹사이트에서 무료로 다운로드가 가능하다(http://pngu.mgh.harvard.edu/~purcell/whap/). TPH2 유전자의 결과를 보면 특정 haplotype 이 ADHD 환자자녀에게 부모로부터 206개가 전이(transmitted)되고 대조군(psedu-control)에게는 151개만이 전이(transmitted)되므로 통계적으로 유의하게 특정 haplotype 이 ADHD 환자자녀에게 전이된 것을 보여준다(p-value =0.004). 또한 연구팀은 600,000 tag SNP genome-wide association scan을 하여 자료분석 중에 있다.

3. The Victorian Family Heart Study (VFHS)

1991년부터 1996년 사이에 2,911명의 건강한 성인 남성들이 VFHS 가족 연구에 포함되었다. 이들 2,911명은 두 명의 부모자료(40-70세 사이)와 적어도 한 명의 자녀 (18-30세 사이) 자료가 포함된 767 가족으로 구성되었다. 연구대상자는 백인(Caucasian)으로 제한되었다. 심장질환 병력이 있는 대상자도 연구에서 제외하였다.

VFHS 연구에서는 현재까지 Genome-wide association 분석은 이루어지지 않았고, 최근에 연구대상자의 키(height)에 대한 Genome-wide linkage 분석을 하여서 3번 염색체에서 키와 관련성이 높은 유전자 위치를 보고하였다 (LOD-score 3.14)[5]. 이 연구에서 연구팀은 400개의 microsatellite marker를 이용하여 전체 22개의 상동염색체와 X 염색체에 걸쳐서 10cM의 해상력(resolution)을 가지는 genome-wide linkage 분석을 하였다. 연구팀은 3번 염색체에서 가장 키와 관련성이 높은 유전자가 있음을 보고하였다. 이 결과로부터 관련성을 보인 3번 염색체 부위(region)에 대해서 1-2cM의 해상력(resolution)을 가지는 fine-mapping linkage 분석을 통해서 3번 염색체의 78cM의 위치에서 키와 높은 관련성이 있음을 보고하였다.

4. The Amish Family Diabetes Study

The Amish Family Diabetes 연구는 제2형 당뇨병 관련 유전자를 찾기 위한 목적으로 1995년에 시작되었다. 제 2형 당뇨병 환자와 그들의 가까운 친척들과 배우자 (18세 이상의 성인)들을 연구에 포함하였다. 1995년에서 1997년 사이에 727명으로부터 자료를 수집하였다. 2004년에 연구대상자 중에서 지질(lipid) 자료가 있는 617명을 대상으로 혈청지질(serum lipid level)에 대한linkage 분석을 하였다[6]. 617명으로 구성된 28가족을 연구하였고, 각 가족은 3명부터 69명까지의 가계도 자료를 포함한다. Linkage 분석을 위해서는 373 microsatellite markers를 통해서 9.7 centimorgans의 평균 밀도(average density)를 갖는 genome-wide linkage 분석을 하였다. 주요 결과는 표4와 같다.

Table 4. Multipoint linkage analysis peaks with LOD>=2.0 (p<0.0012) in the Amish Family Diabetes Study

Chromosome location	Distance (cM)	Closet marker(s)	Trait	LOD (P-value)	Positional candidate genes
2p23	32	D2S312/D2S220	LDL-C	2.17 (0.0008)	APOB, LPIN1, ABCG5, ABCG8
3p25	25	D3S1263	LDL-C	2.47 (0.0004)	PPARG
11q23	135	D11S1345	LNTG	2.03 (0.001)	APOC3, APOA1, APOA4, APOA5
19p13	27	D19S221	LDL-C	2.15 (0.0008)	LDLR
	39	D19S433		2.23 (0.0007)	

LDL-C: serum low density lipoprotein cholesterol, LNTG: ln-transformed TG
Candidate genes: ABCG5: ATP-binding cassette, subfamily G, member5; ABCG8: ATP-binding cassette, subfamily G, member8; APOA1: apolipoprotein A1; APOA4: apolipoprotein A4; APOA5: apolipoprotein A5; APOB: apolipoprotein B; APOC3: apolipoprotein C3; LDLR: low density lipoprotein receptor; PPARG: peroxisome proliferators-activated receptor-gamma
Source: Pollin et al.[6]

즉 LDL 콜레스테롤과는 2p23, 3p25, 19p13 지역에서 높은 LOD-score 값을 보였다. 또한 최근에는 동일 대상자에 대하여 whole-genome wide association 자료에 대한 분석이 진행 중이다.

5. The NIMH Alzheimer's disease Study

1991년에서 1997년 사이에 1,439명의 대상자가 연구에 포함되었다. 적어도 2명의 알츠하이머 환자가 포함된 437 가족이 연구에 포함되었다. 대상자중994명은 알츠하이머 환자, 411명은 정상인이었다. 전체 연구대상자에 대하여 10년간의 추적기간 동안 알츠하이머에 대한 추가진단이 이루어졌다. 2005년에 연구팀은 이미 linkage 분석에서 알츠하이머와 높은 관련성을 보인 염색체 9q22 지역에 3개 유전자 19개 SNP에 대한 관련성 분석을 하였다. 분석은 FBAT 검정과 conditional logistic regression 분석을 하였다. 연구팀은 UBQLN1 유전자에 위치한 SNP에서 알츠하이머와 통계적으로 유의한 관련성을 보고하였다[7].

6. The Breast Cancer Family Registry

The Breast Cancer Family Registry는 미국, 호주, 캐나다의 3개국에서 2003년 9월까지 11,950 가족을 등록하였다. 6,126 population-based case 가족, 2,990 population-based control 가족, 1,647 clinic-based control 가족, 1,187 clinic-based control 가족 등을 포함하였다[8]. 2006년에 연구팀은 ATM유전자와 유방암과의 관련성을 보고하였다. 이 연구에서 관련성은 유방암에 걸린 적이 있는 어머니를 가졌던 환자의 경우 odds ratio=5.5로 그렇지 않은 경우보다 높은 관련성을 보였다[12].

가족기반 코호트 연구에서 주로 사용되는 통계분석방법

1. Linkage 분석

가족기반 코호트 자료를 이용하여서는 우선 linkage 분석을 할 수 있다. Linkage 분석은 질병의 원인이 되는 염색체의 근접한 위치(disease locus)를 재조합(recombination) 현상을 이용하여 찾는 분석방법이다. 재조합(recombination)이란 감수분열 시에 아버지의 염색체와 어머니의 염색체가 교차(crossing-over)가 일어나므로 염색체의 구성이 바뀌는 현상이다. 즉, Linkage 분석에서 다른 염색체에 위치한 유전자들은 감수분열 시 독립적으로 분리(segregate) 되므로 두 유전자 사이에는 아무런 연관이(no linkage) 없다. 그러나, 동일한 염색체 상의 두 개의 유전자 간에는 특정 재조합(recombination) 확률(Θ)을 갖게 되는데 이 확률은 두 유전자의 위치가 멀수록 높아진다. 가장 높은 재조합 확률은 50%로서 이는 다른 염색체에 위치한 유전자들의 재조합(recombination) 확률과 같다. 따라서 특정 유전자 마커(marker)가 질병유전자(disease gene)와 매우 낮은 재조합율(recombination rate)을 갖는다면 질병유전자가 이 마커와 매우 가까운 위치에 있다고 추

정할 수 있다. 이러한 원리를 이용하여서 질병 유전자의 위치를 찾는 것이 linkage 분석 방법이다[13].

일반적으로 사용되는 marker locus로서는 restriction fragment length polymorphisms(RFLP), variable number of tandem repeats(VNTR), microsatellite (예: CA repeats), single nucleotide polymorphism(SNP) 등이 이용 가능하나 최근에는 microsatellite와 SNPs 을 주로 이용한다. VFHS 연구의 예처럼 400여개의 microsatellite를 이용하면 10cM의 해상도(resolution)을 가지는 whole genome linkage 분석을 하는 것이 가능하다. 또한, VFHS 연구의 예처럼 관련성을 보인 염색체 위치에 대해서 추가적으로 보다 높은 해상도(resolution)을 가지는fine-mapping linkage 분석을 할 수 있다.

Linkage 분석을 위해서는 가족 자료가 필수적이다. 가족수가 많은 가계도(extended families)가 보다 효율적이지만 소규모 가족(nuclear families)이나 형 제쌍 자료 (pairs of sibs)도 linkage 분석이 가능하다. 또한 linkage 분석은 크게 모수적 (parametric) linkage 분석과 비모수적 (non-parametric) linkage 분석으로 나뉜다. Linkage 분석을 위해서 최근 많이 이용되는 소프트웨어로는 Merlin 등이 있다. Merlin은 미국의 미시건 대학 web-site에서 무료로 다운로드가 가능하다 (http://www.sph.umich.edu/csg/abecasis/Merlin/).

2. TDT 분석과 FBAT

가족자료 관련성(association) 연구를 위해서는 단순한 case-parents trio 자료를 이용하는 경우는Transmission Disequilibrium Test (TDT) 분석이 이용되고 일반적인 가족자료(extended family-based association study data)에 대해서는 Family based association test (FBAT) package가 많이 이용된다. TDT 분석은 가장 단순한 가족의 형태인 trios (부모, 질병에 이환된 자녀 한명) 자료의 경우에 이용되는 방법이다. TDT 분석방법의 아이디어는 간단하다. 멘델의 법칙을 따르는 귀무가설에서는 특정 marker allele이 환자자녀에게 전이(transmitted)될 확률은 50%이다. 즉 그림1의 general model에서 아버지로부터 환자자녀에게 M1 대립유전자(allele)가 전이(transmitted)될 수도 있고 M2 대립유전자(allele)가 전이(transmitted)될 수도 있다. 즉, M1 대립유전자가 아버지로부터 전이될 확률은 50%이다. 그림의 예에서는 M1 allele을 질병관련 대립유전자(risk allele)로 가정하고 분석을 한다(그림 1). 그림의 예제(example) 1에서는 아버지(dad)로부터 환자에게 M1 allele이 전이(transmitted)되었고, M1이 아닌 allele (M2 allele)이 아버지로부터 전이(transmitted)되지 않았다. 따라서 b cell에 포함(count)된다. 모(mom)로부터는 M1이 아닌 allele(M3)이 전이되었고, 역시 M1이 아닌 allele(M4)이 전이되지 않았다. 따라서 d cell에 포함(count)된다. 이와 동일한 방법으로 모든 trios 자료에 대한 결과를 a,b,c,d cell에 포함(count)하고 이 값을 모두 더한다. 귀무가설 하에서는M1이 전이될 확률은 50%이어야 한다. 따라서 더해진 b cell과 c cell은

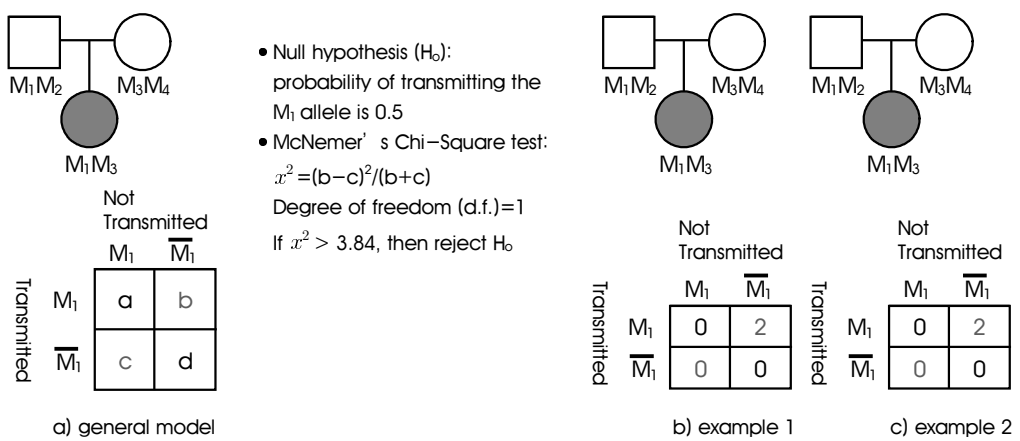

Fig. 1. Example of TDT analysis

귀무가설하에서는 동일하여야 한다. 이것을 통계적으로 검정하기 위해서는 다음의 McNemar의 chi-square test: $X^2=(b-c)^2/(b+c)^2$를 사용할 수 있다[14]. FBAT 분석의 경우는 TDT와 동일한 개념에서 시작하여서 trios 이외에 다양한 가족 자료(extended family-based association study data)에 이용할 수 있도록 개발된 방법이다. 또한 TDT 분석이 이분형 변수(binary trait)에만 적용이 가능한 반면 FBAT은 연속형 변수(quantitative trait)에도 적용이 가능하다. 또한 결측값(missing) 자료가 포함된 가족 자료를 이용하는 것도 FBAT에서는 가능하다. FBAT 소프트웨어는 하버드대학에서 개발한 것으로 하버드 대학 website (http://www.biostat.harvard.edu/~fbat/fbat.htm)에서 무료로 다운로드받을 수 있다[15].

최근에는 Illumina나 Affymerix 등이 개발한 Gene Chip을 이용한 Whole-Genome association study가 많이 이루어지고 있다. 이 자료를 위한 분석방법은 TDT 분석과 동일하나 대규모 자료 분석을 위해 PLINK 소프트웨어가 주로 이용된다[16].

3. Imprinting (parent-of-origin) effect 분석

또한 Imprinting(parent-of-origin) effect를 보기 위한 연구방법으로는 우선 screening 단계에서는 Cordell 등이STATA 프로그램 기반으로 개발한 방법을 이용할 수 있다. Cordell 방법에서는 모성전이(maternal transmission)과 부성전이(paternal transmission)에서 TDT 결과를 단순히 따로 보는 것이 가능하다[17]. 표 5는 Cordell 방법을 이용하여 imprinting effect를 연구한 예이다. 이 연구는 존스홉킨스병원 정신과 연구팀이 bipolar disorder 환자가 있는 344 가족을 대상으로 TDT 분석을 한 것이다. 표 5에서 SNP rs789024에 대해서 전체 부모로부터의 전이(transmission)를 고려한 분석에서는 p-value 0.003의 결과를 얻었고, 아버지로부터의 전이(transmission)만을 고려한 분석에서는 p-value 0.00007로 통계적으로 유의하였으나, 어머니로부터의 전이(transmission)만을 고려한 분석에서는 p-value 0.78로 관련성이 없었다. 따라서 이 SNP은 아버지를 통해서만 bipoloar disorder와 관련이 있다고 추정할 수 있다 [18]. 또한, PLINK 소프트웨어서는 대규모 SNP 자료(GWA 연구)에 대해서도 Imprinting effect를 screening할 수 있는 명령어를 제공하고 있다 [16]. 보다 정확하게 imprinting effect를 보기 위한 방법으로는 Weinberg 등이 개발한 log-linear 모델이 주로 이용된다. 이 방법의 장점은 단순히 imprinting의 효과만 보는 것이 아니라 자녀의 유전자형 결과(genotype effect)와 어머니의 유전자형 결과(genotype effect)를 보정(adjustment)하여 imprinting 결과를 제시한다는 점이다[19].

가족기반 코호트 연구의 장점, 단점 및 문제점

1. 장점

가족 기반 코호트 연구는 유전연구에 있어서 여러 가지 측면에서 주요한 역할을 할 수 있다. 첫째, population stratification의 영향을 적게 받는다. Population stratification이라는 것은 질환과 유전자와의 관련성에 의해서가 아니라 환자군과 대조군이 단순히 다른 조상에서부터 옴으로 인해서 대립유전자 빈도(allele freq-

Table 5. Transmission of alleles to bipolar I offspring, stratified by parent-of-origin of alleles

	Rs789024				Rs1893157		
	T	U	P-value		T	U	P-value
Stratified TDT							
Paternal	77	35	0.00007	Paternal	52	19	0.00009
Maternal	60	57	0.78	Maternal	24	23	0.88
Total	137	92	0.003	Total	76	42	0.002

Source: Mulle et al.[18]
T: transmitted, U: untransmitted

uency)가 차이가 나는 경우로 환자-대조군 연구에서 위양성(false positive) 결과의 원인이 될 수 있다.

둘째, 가족연구에서 통계적으로 유의한 결과는 linkage와 association의 두 가지 의미를 동시에 갖는다. 이것은 linkage 분석이나 association 분석을 따로 하는 것보다 높은 검정력을 갖는다 [2].

셋째, 가족 연구를 통해서는 유전율(heritability)을 추정하는 것이 가능하다.

넷째, 가족연구의 고전적인 장점으로서 familial risk의 평가가 가능하다. 또한 epigenetic study에서 비가족 연구(unrelated individual)로는 얻을 수 없는 정보(information)을 얻을 수 있다. 즉, 모계 유전자, 자녀 유전자, 그리고 imprinting의 효과 등을 분리하여 분석이 가능하다. Imprinting을 본 연구의 예로는 트리오 자료를 이용하여 소아비만에 대하여 imprinting의 효과를 본 연구[20]와 기타 구순구개열, 천식 등 여러 질환에 대해서 imprinting 효과가 연구되고 있다.

2. 단점 및 문제점

가족 기반 연구의 가장 큰 단점은 환자-대조군 연구나 일반적인 코호트 연구에 비하여 대규모 연구대상자를 모으는데 시간과 비용이 많이 든다는 점이다[21]. 또한, 환자-대조군 연구나 일반인구 코호트 연구에 비하여 자료를 분석하는 소프트웨어의 상용화와 보급이 느리다[15]. 또한, 다른 단점으로는 가족연구에서 많이 사용하는 TDT 분석방법의 경우 환경 및 생활 습관 요인의 main effect를 볼 수 없다는 것이다. 그러나, 소프트웨어의 개발과 보급은 지속적으로 이루어지고 있고, 환경 및 생활 습관 요인의 main effect를 함께 연구하기 위한 연구방법론의 개발도 여러 연구자들에 의하여 활발히 논의되고 있다[22]. 즉, 가족을 대상으로 한 분석방법에서 환경적인 요인과 질환의 관련성을 유전적인 요인을 "가족관계의 비독립성을 보정"하는 방법(GEE 등)을 통해서 수행하고 있으며, 두 가지 환경요인간의genetic correlation과environmental correlation을 구분할 수 있는 장점도 가족연구는 가지고 있다.

가족기반 코호트 연구의 방향 및 전망

최근의 유전연구에서Linkage 분석이나 후보 유전자(candidate gene) 연구가 여전히 많이 이루어지고 있고 그 가치가 여전히 인정되고 있으나, 주요연구 흐름은 genome-wide association (GWA) 연구라 하여도 지나치지 않다. GWA 연구에서 가장 큰 문제는 수십만개의 SNP을 검정하는 다중비교(multiple testing) 문제이다. 환자-대조군 연구에서도 이 문제를 해결하기 위하여 주로 multi-stage 디자인을 이용한다. 가족 연구에서도 이 multi-stage 디자인이 가능한데, 이 때 가족 연구의 장점은 환자-대조군 연구와 달리 동일한 data set을 screening 단계와 testing 단계에서 이용할 수 있다는 것이다[23].

GWA 연구를 위해서는 대규모의 연구대상자가 필요하다는 점에는 이견이 없다. 그런 측면에서 부모나 친척의 자료를 모을 필요가 없는 환자-대조군 연구와 일반인구 코호트 연구가 가족 연구에 비하여 대상자를 모으는데 장점이 있는 것이 사실이다. 그러나, 환자-대조군 연구와 일반적인 코호트 연구는 측정되지 않은 population substructure 문제에서 자유로울 수 없다. 그 예로 영국의 Wellcome Trust Case Control Consortium에서 7개 질환에 대한 Genome-wide association 환자 대조군 연구결과를 발표한 논문에서도 위양성의 가능한 원인으로 population structure를 제한점으로 인정하고 있다[24]. 따라서 GWA 연구를 위해서도 여전히 가족-기반 코호트 자료의 구축은 필요하다[15]. 또한 최근 주요한 연구분야가 되는 imprinting에 대한 분석을 하는 것도 가족연구에서만 가능하다. 또한, 자료를 수집하는데 많은 어려움이 있음에도 프램험 연구와 IMAGE project 등은 이미 가족기반 GWA 연구를 진행하여 그 결과를 제시하고 있다. 국내에서도 GWA 연구를 위한 코호트 자료가 질병관리본부의 주도하에 이루어지고 있음은 바람직하다[25]. 본 연구를 통해 GWA를 위한 국내 코호트 구축에 있어서 가족 기반 코호트 구축에 대한 논의도 보다 활발히 이루어지기를 기대한다.

요 약

가족기반 코호트 연구를 하는 것은 단면연구나 일반인을 대상으로 하는 코호트 연구에 비하여 비용과 시간이 많이 걸리는 단점이 있으나 만성질환의 환경요

인과 유전요인을 동시에 연구함에 있어서 많은 장점을 가지므로 연구계획 시 고려할 가치가 있겠다. 그 주된 장점으로는 population structure의 영향을 적게 받는다는 것, imprinting의 효과를 볼 수 있다는 것과 연관성(association) 분석 이외에 linkage 분석도 동시에 할 수 있다는 것이다. 따라서 실제로 미국의 프램험 연구와 유럽의 IMAGE project 등에서 가족기반 코호트 연구를 활발히 진행 중이며genome-wide association 연구를 포함한 성공적인 여러 관련 결과들을 발표하고 있다.

가족 기반 코호트에서 주로 사용되는 연구방법은 linkage 분석과 TDT 분석, 그리고 imprinting effect 분석이다. 이와 관련된 내용은 본 연구에서 간단히 기술하였다. 최근 가족기반 코호트 연구의 방향은 다른 환자-대조군 연구나 일반인구 코호트 유전 연구와 마찬가지로genome-wide association 연구를 하는 것이다. 미국과 영국을 포함한 여러 나라에서 경쟁적으로 가족기반코호트를 이용한genome-wide association 연구를 진행하고 있으나 생활습관요인과 유전요인이 상이한 자국민에 대한 연구는 필수적이다. 또한 genome-wide association 연구를 통한 genome-wide imprinting 연구는 국제적으로 초기단계라 할 수 있다. 따라서 국내에서도 국가적인 지원 하에 가족기반 코호트 연구를 시작한다면 국제적으로 경쟁력 있는 연구를 하는 것이 가능하겠다.

참고문헌

1. Collins FS. The case for a US prospective cohort study of genes and environment. Nature. 2004 May 27;429(6990):475-7.
2. Gauderman WJ, Conti DV. Commentary: Models for longitudinal family data. Int J of Epidemiology 2005;34:1077-1079.
3. Cupples LA, Arruda HT, Benjamin EJ, D'Agostino RB Sr, Demissie S, DeStefano AL, Dupuis J, Falls KM, Fox CS, Gottlieb DJ, et al. The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. BMC Med Genet. 2007;8 Suppl 1:S1.
4. Kuntsi J, Neale BM, Chen W, Faraone SV, Asherson P. The IMAGE project: methodological issues for the molecular genetic analysis of ADHD. Behav Brain Funct. 2006 Aug 3;2:27.
5. Ellis JA, Scurrah KJ, Duncan AE, Lamantia A, Byrnes GB, Harrap SB. Comprehensive multi-stage linkage analyses identify a locus for adult height on chromosome 3p in a healthy Caucasian population. Hum Genet. 2007 Apr;121(2):213-22.
6. Pollin TI, Hsueh WC, Steinle NI, Snitker S, Shuldiner AR, Mitchell BD. A genome-wide scan of serum lipid levels in the Old Order Amish. Atherosclerosis. 2004 Mar;173(1):89-96.
7. Bertram L, Hiltunen M, Parkinson M, Ingelsson M, Lange C, Ramasamy K, Mullin K, Menon R, Sampson AJ, Hsiao MY, Elliott KJ, Velicelebi G, Moscarillo T, Hyman BT, Wagner SL, Becker KD, Blacker D, Tanzi RE. Family-based association between Alzheimer's disease and variants in UBQ LN1. N Engl J Med. 2005 Mar 3;352(9):884-94.
8. John EM, Hopper JL, Beck JC, Knight JA, Neuhausen SL, Senie RT, Ziogas A, Andrulis IL, Anton-Culver H, et al. The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. Breast Cancer Res. 2004;6(4):R375-89.
9. Shearman AM, Cupples LA, Demissie S, Peter I, Schmid CH, Karas RH, Mendelsohn ME, Housman DE, Levy D. Association between estrogen receptor alpha gene variation and cardiovascular disease. JAMA. 2003 Nov 5;290(17):2263-70.
10. Herbert A, GerryNP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, Colditz G, Hinney A, Hebebrand J, Koberwitz K, Zhu X, Cooper R, Ardlie K, Lyon H, Hirschhorn JN, Laird NM, Lenburg ME, Lange C, Christman MF. A common genetic variant is associated with adult and childhood obesity. Science. 2006 Apr 14;312(5771):279-83.
11. Brookes K, Xu X, Chen W, Zhou K, Neale B, Lowe N, Anney R, Franke B, Gill M, Ebstein R,

et al. The analysis of 51 genes in DSM-IV combined type attention deficit hyperactivity disorder: association signals in DRD4, DAT1 and 16 other genes. Mol Psychiatry. 2006 Oct;11(10):934-53.
12. Bernstein JL, Teraoka S, Southey MC, Jenkins MA, Andrulis IL, Knight JA, John EM, Lapinski R, Wolitzer AL, Whittemore AS, West D, Seminara D, Olson ER, Spurdle AB, Chenevix-Trench G, Giles GG, Hopper JL, Concannon P. Population-based estimates of breast cancer risks associated with ATM gene variants c.7271T>G and c.1066-6T>G (IVS10-6T>G) from the Breast Cancer Family Registry. Hum Mutat. 2006 Nov;27(11):1122-8.
13. Thomas DC. Statistical Methods in Genetic Epidemiology. Oxford UniversityPress. 2004; 15-17.
14. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993 Mar;52(3):506-16.
15. Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet. 2006 May;7(5):385-94.
16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep; 81(3):559-75.
17. Cordell HJ, Barratt BJ, Clayton DG. Case/pseudocontrol analysis ingenetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet Epidemiol. 2004 Apr;26(3):167-85.
18. Mulle JG, Fallin MD, Lasseter VK, McGrath JA, Wolyniec PS, Pulver AE. Dense SNP association study for bipolar I disorder on chromosome 18p11 suggests two loci with excess paternal transmission. Mol Psychiatry. 2007 Apr;12(4):367-75.
19. Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Hum Genet. 1999 Jul;65(1):229-35.
20. Le Stunff C, Fallin D, Bougnères P. Paternal transmission of the very common class I INS VNTR alleles predisposes to childhood obesity. Nat Genet. 2001 Sep;29(1):96-9.
21. Hopper JL, Bishop DT, Easton DF. Population-based family studies in genetic epidemiology. Lancet. 2005 Oct 15-21;366(9494):1397-406.
22. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. Am J Hum Genet. 2005 Oct; 77(4):627-36.
23. Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, Demeo DL, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman EK, Laird NM, Weiss ST, Lange C. Genomic screening and replication using the same data set in family-based association testing. Nat Genet. 2005 Jul;37(7):683-91.
24. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 Jun 7;447(7145):661-78.
25. Sung J, Cho SI. Strategy Considerations in Genome Cohort Construction in Korea. Korean J Prev Med. 2007 Mar;40(2):95-101.

=Abstract=

Examples and outlook of family-based cohort study

Jae Woong Sull[1,4], Sue Kyung Park[2], Heechoul Ohrr[3], Sun Ha Jee[1,4]

Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA[1]
Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Korea.[2]
Department of Preventive Medicine and Public Health, Yonsei University College of Medicine, Seoul, Korea[3]
Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea.[4]

Family-based designs are commonly used in genetic association studies to identify and to locate genes that underlie complex diseases. In this paper, we review two examples of genome-wide association studies using family-based cohort studies, including the Framingham Heart Study and International Multi-Center ADHD Genetics Project. We also review statistical methods of family-based designs, including the transmission disequilibrium test (TDT), linkage analysis, and imprinting effect analysis.

In addition, we evaluate the strengths and limitations of the family-based cohort design. Despite the costs and difficulties in carrying out this type of study, a family-based cohort study can play avery important role in genome wide studies. First, the design will be free from biases due to population heterogeneity or stratification. Moreover, family-based designs provide the opportunity to conduct joint tests of linkage and association. Finally, family-based designs also allow access to epigenetic phenomena like imprinting. The family-based cohort design should be given careful consideration in planning new studies for genome-wide strategies.

Key Words: Family-based cohort study, transmission disequilibrium test (TDT), linkage study