예방의학회지 제40권 제2호(2007년 3월)
J Prev Med Public Health 2007;40(2):95-101

특 집

한국 유전체 코호트 구축의 전략적 고려사항

성주헌, 조성일[1)]

강원대학교 의과대학 예방의학교실, 서울대학교 보건대학원 역학교실[1)]

Strategy Considerations in Genome Cohort Construction in Korea

Joohon Sung, Sung-il Cho[1)]

Department of Preventive Medicine, Kangwon National University College of Medicine,
Department of Epidemiology, Seoul National University School of Public Health[1)]

Focusing on complex diseases of public health significance, strategic issues regarding the on-going Korean Genome Cohort were reviewed: target size and diseases, measurements, study design issues, and follow-up strategy of the cohort. Considering the epidemiologic characteristics of Korean population as well as strengths and drawbacks of current research environment, we tried to tailor the experience of other existing cohorts into proposals for this Korean study. Currently 100,000 individuals have been participating the new Genome Cohort in Korea. Target size of *de novo* collection is recommended to be set as between 300,000 to 500,000. This target size would allow acceptable power to detect genetic and environmental factors of moderate effect size and possible interactions between them. Family units and/or special subgroups are recommended to parallel main body of adult individuals to increase the overall efficiency of the study. Given that response rate to the conventional re-contact method may not be satisfactory, successful follow-up is the main key to the achievement of the Korean Genome Cohort. Access to the central database such as National Health Insurance data can provide enormous potential for near-complete case detection. Efforts to build consensus amongst scientists from broad fields and stakeholders are crucial to unleash the centralized database as well as to refine the commitment of this national project.

J Prev Med Public Health 2007;40(2):95-101

Key words : Cohort studies; Human genome; Genetic predisposition to disease; National Health Insurance, Republic of Korea; Follow-up studies; Biological specimen banks

서론

유전체 코호트 구축의 배경

건강-질병 현상과 치료에 대한 반응을 규정하는 유전-환경적 요인을 구명하는 것은 질병부담을 줄이기 위한 연구의 핵심이다. 이는 전체 의과학 연구의 가장 중요한 목표이기도 하다. 인간게놈프로젝트를 완수함으로써, 인류는 새로운 차원에서 유전-환경에 대한 연구를 수행할 수 있게 되었다. 또한, 과학이 답할 수 없었던 질문들, 예를 들어, 개인의 유전적 체질이 생활-환경요인과 결합되어 어떤 질병의 발병위험이 더 높아 지는지 알 수 있을 것이라는 기대도 할 수 있게 되었다 [1].

유전적인 요인의 규명과 환경적인 요인의 규명은 역사적으로 다른 맥락과 방법론 그리고 학자들에 의해 이루어져 왔다. 유전학은 혈우병이나 알캅톤뇨증과 같은 단일 유전자질환(mendelian disease)의 발견으로 촉발되어, 특정 질환을 가진 가계에 대한 연구를 중심으로 이루어 졌다. 유전형 분석이 가능해진 1980년대 후반 이후 지금까지 2,000 개에 이르는 단일 유전질환의 유전자가 발견되었다 [2]. 즉, 유전학은 일반인구 집단이 아닌 특정 유전요인이 고도로 집적된 특수집단(=질환가계)을 선택함으로써, 성공적으로 유전요인을 발견해 왔고, 고위험 유전인자를 가진 사람의 예방전략을 발전시켜 왔다 [3]. 반면, 환경적인 요인의 구명은 전통적인 역학적 연구가 뼈대를 이루어 왔다. 콜레라에 대한 John Snow의 연구를 필두로, 현대에도 흡연과 폐암, 콜레스테롤과 심혈관질환 등 건강증진에 결정적인 역할을 한 성과들이 인구집단에 대한 역학연구를 통해 이루어졌다.

유전역학(genetic epidemiology) 혹은 유전체역학(human genome epidemiology)은 인체 유전학(human genetics)과 전통적인 역학적 방법론이 결합되어 특히 복합질환(complex disease)의 원인구명을 위한 방법론으로 제시되었다. 유전역학 혹은 유전체 역학은, 같은 이름을 걸고 최근까지도 크게 두 가지의 흐름으로 전개되고 있다고 생각된다. 즉, 유전학의 연장선상에서 유전역학을 수행하고자 하는 흐름과 역학, 특히 분자역학적인 기반의 연장선상에서 유전체역학으로 영역을 넓혀나가려는 흐름이 있었고, 각각 성과와 한계를 노정하고 있다. 인체 유전학을 기반으로 한 유전역학 연구들은 복합질환의 원인 규명에서 단일 유전질환과는 다른 도전에 직면하게

이 연구는 질병관리본부 학술용역사업(2006-347-2400-2440-215)으로 수행된 내용에 근거한 것임.
책임저자 : 조성일 (서울시 연건동 28번지, 전화 : 02-740-8868, 팩스 : 02-747-7082, E-mail : scho@snu.ac.kr)

된다. 환경요인이 중요한 역할을 하며, 유전-환경요인의 교호작용이 있을 것으로 생각되는 복합질환에서, 기존의 gene mapping 연구방법들은 일반적으로 부분적인 성공만을 거두었다. 알츠하이머병의 PS1, PS2, APP 유전자, 유방암의 BRCA1, BRCA2 등 복합질환의 유전자들도 일부 특수한 형태의 질환군만을 설명하고 있다 [4]. 한편, 역학의 전통에서 출발한 연구들은 유전적인 요인이 회상 바이어스(recall bias)에 영향을 받지 않는 "과거의 노출요인"이라는 점에 착안하여 환자-대조군 연구(case-control study)를 중심방법론으로 유전요인-질병의 관련성을 검증하고자 하였다 [5]. 이런 환자-대조군 연구는 당초의 많은 기대와는 달리 한계점을 드러내었다. 즉, 환자대조군 연구 일반이 가질 수 있는 환자의 선택 바이어스(selection bias)나 생환-환경요인들에 대한 회상 바이어스의 문제들은, 33억 개에 이르는 염기서열들 모두가 잠재적인 후보가 될 수 있다는 전혀 새로운 차원의 다중비교(multiple comparison)와 결부되어 심각한 위양성(false-positive), 출판 바이어스의 문제가 제기되었다. 최근 십년 동안 매년 수 천 편이 넘게 발표되는 연구결과들 중에서 결과가 재현되고 공인되는 성과는 손에 꼽힐 정도가 되었다 [3]. 이론상으로 5%의 위양성을 허용하다면, 33억 개의 염기서열 중 5% 즉, 1억 6천만 개의 위양성 결과가 가능하다. 기존의 역학연구에서 환경요인에 대한 평가의 부정확성은 단순오류로 종결될 수 있었지만, 유전적 요인과의 교호작용을 다루기 시작할 때는 천문학적인 위양성의 가능성을 만들어내고, 출판 바이어스 등으로 귀결될 수 있다. 단순한 다중비교의 보정인 Bonfe- rroni 방법 등으로 이를 보정하려고 하면 어떤 연구결과도 유의한 결과를 얻기 힘든 문제가 발생한다. 최근에는 유전형 분석기법의 비약적인 발달로 전장유전체 분석(genome-wide analysis) 방법과 tagSNP 등을 이용하여 후보 유전자 연구(candidate g- ene study)가 가지는 문제는 많이 해결될 수 있지만 [6], 여전히 환자-대조군 연구 일반의 한계와 다중비교의 문제는 숙제로 남아 있다.

최근 가장 중요한 질환으로 부각되고 있는 심혈관계질환, 당뇨, 일부 악성종양 등은 유전적 감수성을 가진 사람들이 급격한 생활환경의 변화를 겪으면서 발생한 현대의 "유행병(epidemic)" 이며, 따라서 이러한 원인들을 다루기 위해서는 유전적인 요인과 환경적인 요인 모두에 대한 종합적인 접근이 필수적이라는 데에 별다른 이견은 없는 것으로 보인다 [1]. 흡연을 제치고 공중보건의 주적(主敵)으로 부상하고 있는 비만의 문제는 대표적으로 이러한 범주에 해당된다. 기존의 유전역학적인 연구의 한계점에 대한 반성과 대안으로, 몇몇 나라에서는 대규모의 투자와 시간, 인력을 요구하는 야심적인 "유전체 코호트" 추진을 결정하였다. 예비연구 단계를 거쳐 올해 본격적으로 시작되는 영국의 UK Biobank [7], 일본의 Biobank Japan [8], Iceland 인구의 30%에 이르는 사람들의 유전체와 가계도, 질환정보를 구축한 deCODE사의 Iceland Biobank [9] 등등이 대표적인 사례이며, 에스토니아, 독일, 캐나다에서도 시작이 되고 있고, 미국을 비롯한 많은 나라들에서 독자적인 유전체 코호트의 구축을 시작하고 있거나 논의되고 있다 [1]. 우리나라에서도 이미 2004년도부터 질병관리본부가 중심이 되어 학계와 공동으로 유전체 코호트 구축사업을 진행하고 있다. 기존에 축적된 경험을 바탕으로 연구설계가 이루어졌고, 현재 약 10만 여명에 이르는 참여자를 확보한 상황이다. 그러나 정확한 목표규모와 대상질환, 다양한 세부 구성요인들 간의 표준화와 통일성 확보, 윤리적인 검증의 문제, 연구에 참여한 사람들이 제공한 귀중한 정보들을 어떻게 질환발생 여부와 성공적으로 결합시킬 것인가 등의 문제는 아직 해결되지 못한 주요 과제로 남아 있다.

이 연구에서는 지금까지의 유전체 코호트에 대한 논의들에 대한 고찰을 통해서, 우리나라에서 새롭게 추진되고 있는 유전체 코호트의 전략적인 고려사항들을 점검해보고 코호트의 목표와 과제들에 대한 나름대로의 제안을 하고자 한다. 우리나라가 고유한 유전적-환경요인을 가지기 때문에 독자적인 연구가 필요하다는 당위의 차원을 넘어서서, 국제적인 과학의 발전과정 속에서 우리나라의 유전체 코호트

Table 1. Estimated disease prevalence (per 1,000 persons per year), incidence (per 100,000 persons per year) and mortality (per 100,000 persons per year) rates of major disease groups according to the disease classification of global burden of disease study

Disease group	Prevalence		Incidence		Mortality	
	Men	Women	Men	Women	Men	Women
Stomach cancer	1.2	0.7	48.1	26.0	23.5	18.5
Colon and rectum cancers	0.6	0.5	20.9	14.6	9.1	8.2
Liver cancer	0.8	0.2	34.9	10.1	31	8.6
Pancreas cancer	0.1	0.1	6.1	4.0	8.3	5.9
Trachea, bronchus and lung cancers	0.6	0.2	31.7	11.2	29	13.5
Breast cancer	0.0	0.9	0.1	25.8	0	3.3
Cervix uteri cancer	0	0.6	0	17.5	0	4.6
Prostate cancer	0.1	0	4.6	0	3.5	0
Lymphoma and multiple myeloma	0.1	0.1	5.2	3.3	3	2.3
Leukaemia	0.1	0.1	3.9	2.8	3.3	3
Alcohol abuse	0.9	0.1	37.1	3.7	4.1	0.5
Schizophrenia	2.2	1.9	48.9	41.0	0.2	0.2
Unipolar major depression	1.3	2.8	49.3	95.4	0	0
Bipolar disorder	0.4	0.4	10.7	11.0	0	0
Dementia and degenerative CNS disorder	0.3	0.5	13.0	22.6	6.6	22.3
Epilepsy	2.4	1.8	57.9	42.9	0.7	0.5
Parkinson disease	0.3	0.4	6.8	9.7	1.3	0.8
Glaucoma	1.1	1.2	25.6	25.0	0	0
Cataracts	1.5	2.4	53.9	73.4	0	0
Diabetes mellitus	11.2	11.2	181.3	148.9	22.2	25.5
Ischemic heart disease	2.3	1.7	63.6	45.8	24.7	22.4
CVA (cerebrovascular attack)	3.6	3.5	125.7	133.5	61.7	86.9
COPD (chronic obstructive pulmonary disorder)	7.5	7.1	215.8	209.5	12.4	9.2
Asthma	10.9	9.3	272.0	243.9	6.3	10.5
Peptic ulcer disease	16.5	19.5	470.6	457.1	2.3	1
Cirrhosis of the liver	3.6	0.8	102.6	20.1	32.2	7.7
Reumatoid arthritis	0.7	2.8	18.4	62.1	0.3	1
Osteoarthritis	2.8	11.5	74.2	225.4	0	0.2
Congenital Anomalies	0.5	0.3	28.3	15.6	0.3	0.2

Table 2. Estimated number of cases and power to detect the smallest effect size(as relative risk, RR), according to the different target cohort size, by the incidence rates, etiologic factors (genetic or environmental), and follow-up duration (5 year and 10 year follow-up). adapted from Manolio et al

IR*	Disease examples		Cohort size of 200,000								Cohort size of 500,000								Cohort size of 1,000,000							
			# of cases		Detectable Effect Size (RR)						# of cases		Detectable Effect Size (RR)						# of cases		Detectable Effect Size (RR)					
					G†		E†		GxE†				G†		E†		GxE†				G†		E†		GxE†	
			Follow-up duration (years)								Follow-up duration (years)								Follow-up duration (years)							
			5	10‡	5	10	5	10	5	10	5	10‡	5	10	5	10	5	10	5	10‡	5	10	5	10	5	10
10	Parkinson disease, schizophrenia	Bipolar disorder	91	201	4.0	2.7	5.0	3.3	>10	>10	228	504	2.5	1.8	3.0	2.3	>10	6.5	457	1,010	1.8	1.7	2.5	1.8	6.7	3.7
50	Colorectal cancer, renal failure	Stomach cancer, myocardial infarction	456	1,008	1.9	1.9	2.5	2.0	7.0	4.0	1,141	2,522	1.7	1.4	1.8	1.7	3.5	2.4	2,282	5,043	1.5	**1.3**	**1.5**	**1.4**	2.5	2.0
100	Breast cancer, hip fracture	Stroke	912	2,016	1.5	1.4	1.8	1.7	4.0	2.5	2,279	5,037	**1.3**	**1.3**	1.6	**1.4**	2.5	**2.0**	4,559	10,075	**1.3**	**1.3**	**1.4**	**1.3**	**2.0**	**1.8**
200	Diabetes, stroke, heart failure	Diabetes, asthma	1,820	4,022	1.4	1.3	1.6	**1.4**	3.0	2.2	4,550	10,056	**1.3**	**1.2**	**1.4**	**1.3**	**2.0**	**1.8**	9,100	20,111	**1.2**	**1.2**	**1.3**	**1.2**	**1.8**	**1.6**
500	Myocardial infarction, all cancers	All cancers, peptic ulcers	4,524	9,998	**1.3**	**1.3**	**1.3**	**1.3**	**2.0**	**2.0**	11,309	24,993	**1.2**	**1.1**	**1.2**	**1.2**	**1.7**	**1.5**	22,618	49,986	**1.1**	**1.1**	**1.2**	**1.1**	**1.5**	**1.3**
3,000	Cataracts, hypertension	Hypertension	25,858	57,146	**1.1**	**1.1**	**1.2**	**1.2**	**1.5**	**1.7**	64,644	142,863	**1.1**	**1.1**	**1.1**	**1.1**	**1.3**	**1.2**	129,289	285,729	**1.1**	**1.1**	**1.1**	**1.1**	**1.3**	**1.1**

* IR - incidence rate per 100,000 person per year † G-genetic effect, E-environmental effect, GxE- gene by environmental interaction effect ‡ aging effect was also considered in calculating the number of cases Bold figure denotes for acceptable level of detection power, here effect size of 1.3, 1.5 and 2.0 were chosen for genetic, environmental and gene-environmental interaction, respectively

가 기여할 수 있는 장점은 무엇이며 이를 어떻게 극대화할 수 있을 것인가 하는 문제의식을 중심으로 기술하고자 하였다.

본 론

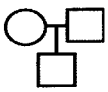
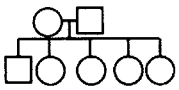
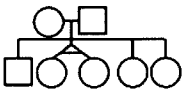
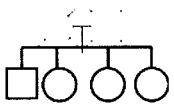
1. 유전체 코호트의 주요 대상질환과 필요한 규모

1990년대 중반 이후에 악성종양은 심혈관계질환을 제치고 우리나라에서 가장 큰 질병부담을 주는 질환이 되었고, 이러한 경향은 더욱 강화되고 있다 [10]. 질병부담에 대한 여러 가지 평가 방법 중에서 가장 널리 사용되고 있는 방법 중의 하나는 세계보건기구에서 제안된 총체적질병부담(Global Burden of Disease, GBD) 방법이며, 이것은 조기에 사망함으로써 생기는 조기사망손실과 질병에 이환되어 발생하는 장애에 의한 손실을 시간으로 합산하여 사망과 상병을 종합한 질병부담을 평가하는 방법이다 [11]. 사고 및 손상을 제외할 때, 2002년도에 우리나라의 질병부담을 평가하면, 전체 악성종양을 필두로 심혈관계질환, 소화기계질환, 당뇨 등이 그 뒤를 잇고 있다. 즉, 우리나라에서 구축되는 유전체 코호트는 우리나라에서 가장 큰 질병부담을 주고 있는 악성종양, 심혈관계질환, 소화기질환, 당뇨, 호흡기질환, 퇴행성 신경계 및 근골격계질환 등의 원인 규명이 가능하도록 설계되고 추진되어야 한다는 것이다.

이러한 주요한 질환의 발생률, 유병률, 사망률 등 주요 빈도지표들은 Table 1에 제시하였다. 질병의 빈도는 약 100 만 명의 표본을 우리나라 전체 인구에 대해서 성별, 연령별, 지역별 대표성을 갖도록 무작위 층화추출 한 후, 이들의 질병발생 및 사망양상을(1998년 ~ 2002년 사이) 의료보험 청구자료와 암등록자료, 사망원인자료 등의 가용한 자료원들을 토대로 1차로 검색하였다. 1차로 검색된 결과들은 환례 기준을 작성하여 해당 기준에 부합되는 사람만을 환자로 포함하였다. 예를 들어, 뇌졸중의 경우는 입원의 기록이 있고 입원기간이 2일 이상이면서 2회 이상 동일한 진단이 4년 안에 있을 경우를 뇌졸중으로 인정하였다. 이렇게 산출된 질환의 빈도지표들은 다시 2,000 명 정도의 표본에 대한 실제 의무기록 조사를 통해 빈도를 보정하여 최종적으로 표 1의 결과를 얻었다. 우리나라의 주요 질병부담을 구성하고 있는 질환들은 10만 명당 발생률이 30 명에서 200명 사이인 경우가 대부분임을 알 수 있다 [12].

이러한 질환들이 모두 유전적, 환경적 요인들과 함께, 유전-환경, 유전자간의 상호작용을 통해 질환의 위험이 결정된다고 하면, 몇 가지의 가정을 통해서 필요한 코호트의 규모를 예상해 볼 수 있다. UK Biobank는 동일한 문제의식을 가지고 약 50만 명 규모의 인구집단을 구축하여 약 20년간 추구관찰을 하는 것이 필요할 것으로 추정하였다 [7]. 또한, 미국의 국립인간유전체연구소(National Human Genome Research Institute) 에서는 위험요인을 유전적 위험요인, 환경적 위험요인, 유전-환경의 상호작용, 유전간의 상호작용 등으로 나누어 평가하였으며, 5년간의 추구조사를 통해서 파악할 수 있는 요인들은 10만 명 당 50 명 정도의 발생률을 보이는 질환

Table 3. Type of recruited families and sampling efficiency for typical genetic analyses

Family unit	Contributions to each analysis				Sampling efficiency (per person)			Remark
	Pedigree examples	Classic Linkage	IBD Linkage	TDT	Classic Linkage	IBD Linkage	TDT	
Family History	?	0	0	0	-	-	-	Cheap, but dependent on the size of family
Affected relative pairs (sibs or other)		0	1	0	0	$<\frac{1}{2}$	0	Suitable for detected cases in cohort
Trio		0	<2	1	0	$<<\frac{2^{\dagger}}{3}$	$\frac{1}{3}$	Necessary for the study of young onset diseases
Nuclear Pedigree (n=4)		<4	5	2	<1	$\frac{5}{4}$	$\frac{1}{2}$	Linkage phase determination can be often ambiguous for nuclear families
Large Pedigree (n=7)		10	20	5	$\frac{10}{7}$	$\frac{20}{7}$	$\frac{5}{7}$	Most efficient, but most difficult to recruit. often linkage phase can be successfully reconstructed.
Large pedigree with twin (n=7)		8	19	4<	$\frac{8}{7}$	$\frac{19}{7}$	$\frac{4}{7}$	Monozygotic twins only contribute heritability estimation
Incomplete large pedigree (n=4)		<<8	6	<4	$<<\frac{8^{\dagger}}{3}$	$\frac{6}{4}$	$<<\frac{4^{\dagger}}{4}$	Power decreases due to incomplete genotype reconstruction

* Affected relative pairs often suffers from uncertainty of IBD sharing
† Actual power for IBD-based linkage drops more because parent-offspring pair does not have IBD distribution (i.e., fixed to 0.5)
‡ Incomplete pedigree may result in severe loss of power, especially when the marker information is not high.
IBD : Identity by descent TDT : Transmission disequilibrium test

의 경우 백 만 명 규모의 코호트에서는 1.5배 정도의 위험도를 환경요인, 이 십 만 명 규모의 코호트에서는 2.3배 정도의 위험도를 가진 환경요인을 규명할 수 있는 것으로 추정되었다. 유전적인 요인은 우성인자라고 가정할 경우 두 개의 대립유전자(allele) 중 하나만 있어도 찾을 수 있기 때문에, 조금 더 낮은 위험도를 가진 요인도 규명할 수 있는 것으로 추정되고 있다.

Table 2는 Manolio 등의 연구 [13]와 동일한 방법을 적용하였을 때, 우리나라에서 중요한 질환의 발생률별로 파악할 수 있는 환경, 유전, 환경-유전 상호작용의 위험도 수준을 5년 및 10년의 추구관찰 결과를 전제로 제시한 것이다. 10년 추구관찰의 경우에서는 코호트참여자들의 노령화에 따른 발생률의 증가를 고려하여, 전반적으로 10%의 발생률 증가가 추가로 생길 것으로 가정하였다. 환경요인의 경우는 1.5배 이하, 유전적 요인의 경우는 1.3배 이하, 환경-유전 상호작용의 경우는 2.0배 이하의 위험도를 가진 요인들도 찾을 수 있는 검정력을 갖는 것을 목표로 할 때, 50만 명 규모의 코호트에서도 10만 명당 100명 정도의 발생률을 갖는 질환만 평가가 가능할 것으로 추정된다. 이러한 추정을 위해서 환경적인 요인의 경우 코호트 참여자의 약 10% 정도가 노출되어 있으며, 유전적 요인의 경우는 10% 정도의 대립 유전자 빈도(allele frequency)를 가지고 있고, 우성효과를 가지고 있어서, 두 개의 대립유전자 중 하나만 있어도 질환의 위험을 높이는 것으로 가정되었다.

우리나라에서 적어도 10만 명당 약 50명 정도의 발생률을 가지는 질환에 대한 연구를 수행할 수 있어야 한다면, 현재 진행되고 있는 코호트는 그 목표가 최소한 천 만 인*년, 즉 백만 명 규모의 코호트일 경우는 십년의 추구관찰이 진행되는 규모이어야 한다. 즉, 이 십 년의 추구관찰이 이루어 질 수 있다면 약 50만 명 규모의 코호트가, 삼십년의 추구관찰에는 30만 명 규모의 코호트 구축이 필요하게 된다.

2. 기존에 구축된 코호트와 새로운 유전체 코호트

현재 우리나라에서 비슷한 수준의 환경적 요인에 대한 조사가 수행되고, 유전체 및 생체시료들이 확보되고 있는 코호트로는 다기관 암코호트 등이 있으며, 확보된 코호트의 구성원은 현재 이미 9만 명에 이르고 있고 장기적으로는 약 20만 명 수준에 도달할 것으로 기대 된다 [14]. 기존의 코호트와 통합하여 유전체 코호트가 구상될 것인지 혹은 새로운 유전체 코호트만으로 필요한 표본수를 채워나갈 것인지는 특히 미국에서 큰 논쟁이 진행되고 있다 [15,16]. 미국에서는 국가가 주도하는 유전체 코호트의 출발이 영국, 일본 등은 물론 우리나라에 비해서도 늦었으며, 기존에

구축되어 온 대규모의 코호트가 이미 검증되어 있는 상황이기 때문에 거액의 국가적인 투자가 다시 필요한지에 대해서 논란이 이루어지는 것은 당연하다. 하지만, 우리나라의 경우는 기존에 구축된 대규모 코호트가 거의 없고, 새롭게 구축되어가고 있는 유전체 코호트가 이미 성공적으로 진행되고 있다는 점, 또한, 기존의 코호트들도 비슷한 문제의식과 프로토콜을 가지고 출발한 경우가 많기 때문에 상당한 부분의 내용들이 호환될 수 있는 가능성이 있다는 점 등은 미국과 다른 상황이라고 할 수 있다 [17]. 즉, 우리나라의 상황에서는 현재와 같이 유전체 코호트가 중심이 되어 독자적으로 30만 명 정도 혹은 그 이상의 코호트를 확보하는 것이 필요하다고 판단된다. 이것은 다기관 암 코호트, 원전 코호트 등의 다양한 코호트들이 호환가능한 프로토콜을 가지고 진행될 수 있다고 할 때 현실적으로는 약 50만 명 규모의 코호트가 구축되는 것과 같은 효과라고 할 수 있다.

3. 코호트의 모집과 참가자에서 수집되는 정보

유전체 코호트는 프로토콜의 통일성 제고를 위한 별도의 연구과제가 수행되는 등 조사도구와 조사내용, 수집되는 정보의 수준을 제고하고 표준화 하려는 노력이 진행되고 있다. 표준화와 통일성 제고는 코호트의 성패를 좌우하는 매우 중요한 핵심적인 과제이다. 특히 코호트 연구의 특장점인 환경요인의 측정이 최대한 정확하게 그리고 각 단위 간에 통일적으로 이루어 질 수 있도록 많은 노력이 기울여져야 함은 재삼 강조될 필요가 있다. 즉, 환경요인을 질병발생 이전의 시점에서 정확하게 평가하지 못한다면, 거액의 투자를 통해서 코호트를 구축해야 하는 의의가 반감되기 때문이다. 유전요인은 일생을 통하여 거의 변화가 없고 훨씬 더 비용효과적인 환자-대조군 연구에서도 똑같이 평가될 수 있음을 명심해야 한다.

수집되는 정보의 또 다른 축인 생체시료들의 경우는 대규모의 연구에 적합한 체계적인 biobank의 구축과 운영, 관리가 시급히 요구된다. 이러한 biobank의 존재는 향후 유전체학(genomics), 단백체학(proteomics) 등을 통해 미래의 과학기술의 발전에 따라 그 용도가 지속적으로 개발 될 수 있다 [1,6,15]. 따라서, 양질의 생체시료를 확보하기 위한 노력은 극히 중요하며 현재 시안이 마련된 표준 운영지침(Standard Operating Procedure, SOP) 을 확대 강화하여 모든 코호트 구축의 단위 기관들에게 엄격하게 적용되어야 할 것이다.

4. 코호트의 효율적인 목표달성을 위한 특수집단의 선정

코호트의 규모에 관한 문제에서 반드시 짚고 넘어가야 할 문제는 유전적 요인에 관한 가정이 너무도 단순하다는 점이다. Manolio 등의 추정은 "다빈도 유전요인에 의한 다빈도 질환"(common disease common variant, CDCV)의 가설을 전제로 하여, 10% 정도의 빈도를 가진 단일한 유전요인이 질환의 발생 위험을 높이는 것을 기본 전제로 하고 있다 [4]. 그러나, CDCV 가설은 많은 사람들의 기대에도 불구하고, 아직 한 번도 실증적으로 검증된 바가 없으며, 훨씬 더 낮은 빈도를 가진 유전요인들의 집합이 다빈도 질환의 유전적 원인일 가능성도 얼마든지 있다. 또한, 해당 유전요인을 가진 사람에게서 질환이 발현되는 침투율(penetrance)의 개념이 우성효과를 갖는다는(즉 해당 대립유전자가 하나만 있으면 질환의 위험도가 증가하는) 가정으로 단순화 되어있으나, 가산적인 공동우성 형태(co-dominant), 즉 한 개의 대립유전자를 갖는 경우는 위험도가 절반이 되는 경우나, 혹은 열성(recessive)의 형태, 즉, 두 개의 대립유전자를 가져야만 비로소 질환발생의 위험이 증가되는 형태 등 개연성이 있는 다른 가정들은 배제되어 있다. 더구나, 하나의 유전요인이 독자적으로 위험을 높이는 것이 아니라 여러 개의 유전요인이 관여하여 질환의 위험도가 결정되는 것이 기존의 질환의 병태생리에 대한 지식과 부합된다면, 이러한 가정들은 단순하기에 앞서 너무 낙관적인 것이라는 비판을 면할 수 없다. 악성종양과 같은 질환들은 환경적인 발암요인이 체세포성 돌연변이를 일으켜 발생하는 산발적인 발생 환례(sporadic cases) 들의 비중이 매우 클 것으로 예상되며, 이러한 경우는 훨씬 더 많은 환례가 있어야만 동일한 위험도를 가진 유전 요인을 규명할 수 있다는 점도 간과되어서는 안 된다.

이런 다양한 유전적 기전들이 있다는 것 자체가 복합질환에서 예외가 아닌 복합질환 자체의 정의이기 때문에, 유전적인 요인에 대한 접근은 일반인구에 대한 접근만으로는 해결되기 어려운 한계가 있다. 이를 극복하기 위해서는 일반인구 단위의 모집전략과 병행하여 가족단위에 대한 모집전략을 수립할 필요가 있다. 예를 들어, 가족단위의 모집이 병행될 경우, 유전자연관분석(linkage analysis)을 통한 gene mapping 연구를 수행할 수 있기 때문에, 후보 유전자좌의 염색체 상에서의 위치를 제시해 줄 수 있게 된다 [18]. 향후, 전장 유전체(genome-wide) 에서 수십만개의 SNP를 검색하는 방법이 표준적인 분석방법이 된다고 가정하면, 가족 연구를 통해 제시되는 후보 유전자자리의 존재는 최대의 난제 중 하나인 다중비교의 문제를 해결할 수 있는 가장 좋은 도구가 될 것이다.

또한, 환경적인 요인에서 생애주기별로 중요성을 가지는 특정 시기의 노출은 해당 연령 군이나 해당 인구집단에 대한 조사를 통해서 보완될 필요가 있다. 예를 들어, 모태 내 환경은 평생 동안의 건강수준을 좌우할 수 있는 중요한 기간으로 평가되고 있지만, 성인에서는 물론, 학동기 정도의 소아만 되어도 정확한 측정이 거의 불가능한 요인이다 [19]. 생애주기에서의 특수집단 이외에도, 도시와 농촌, 다양한 사회경제적인 계층 등에 대한 고려를 통해서 환경적인 요인의 변이가 충분히 확보될 수 있고, 결과의 일반화가 가능하도록 연구대상의 확보가 고안될 필요가 있다.

우리나라가 단일민족 집단이라는 기존의 고정관념에서 벗어나 현실을 볼 때, 급증하는 국제결혼으로 이미 상당수의 외국출신 한국인이 하나의 소수집단을 이루고 있다. 선진국의 경우는 연령과 성별에 대한 고려만큼이나 다양한 민족적인 배경에 대한 고려가 모든 연구와 정책수립의 필수적인 전제조건이다. 코호트가 최소한 10년에서 20년 후에 연구성과가 나타나는 것이라고 했을 때, 현재 증가하고 있는 외국인 이민자들은 우리 사회의 한 구성원으

로 자리 잡을 것이 확실하며, 역시 코호트에 포함되어야 할 특수집단이 될 것이다.

5. 가족단위 연구의 병행을 위한 고려사항

가족단위의 연구는 복합질환의 연구에서도 몇 가지의 고유한 장점을 가지고 있으며, 이 때문에 미국의 유전체 코호트 구축을 위해서도 중요한 연구과제의 하나로 거론되고 있다. 따라서 가족 연구의 고려사항을 별도의 주제로 간략히 다루어보고자 한다. 가족 연구가 기여할 수 있는 장점은 다음과 같다. 1) 가족적인 집적(familial aggregation)을 보이는 환례는, sporadic case가 아닌, 유전체 연구의 목표인 유전적인 감수성의 공유에 의한 확률이 더 높다. 2) 가족적인 집적을 보이는 환례들은 동일한 유전적 요인이 작용할 확률이 혈연관계가 없는 환례들 사이에서 보다 월등히 높아서 보다 큰 검정력을 가진다. 예를 들어, 한 명의 유방암 환자에서는 A 및 B 라는 유전요인과 고지방 식이, 장기간의 여성 호르몬 노출 등이 질환발생의 원인이 되었고, 또 다른 유방암 환자에서는 C 및 D 라는 유전요인과 흡연, 환경성 발암물질이 질환발생의 원인이 되었다고 가정한다면, 일반인구를 대상으로 한 연구에서는 새로운 환자가 추가되어도 A 및 B라는 유전요인과 C 및 D 라는 유전요인은 기존의 모든 분석에서 상쇄되어 위험요인으로 밝히기 어렵게 된다. 3) 가족적 형태의 연구에서는 기존에 알려진 지식과는 전혀 무관하게 유전자가 위치하고 있는 후보 유전자좌를 제시하는 gene mapping 연구를 수행할 수 있어서, 인구집단 연구와의 상승효과를 기대할 수 있다. 분석방법도 최근 급격한 진화를 보이고 있으며, 복합질환을 위한 방법론들이 속속 개발되고 있다 [20,21]. 4) 일반인구의 전장유전체 분석에 통상 50만 ~ 60만 개정도의 유전형 분석이 필요한 것과는 달리, 가족연구는 400개 ~ 1,000개의 유전형 분석만으로도 충분히 결과가 나올 수 있어서 비용효과적이다.

반면, 가족 단위연구는 몇 가지 단점을 가지고 있어서 1) 일반인구의 개인단위 모집에 비해서 가족 전체를 포괄시키기 위한 노력은 훨씬 더 힘들다. 2) 가족은 가족의 크기에 따라서 검정력에 큰 차이를 가져오게 되어, 예를 들어 100 명의 가족을 모집한 경우라도, 3일가족을 중심으로 30여 가족을 모집한 경우(90 명) 보다 8인 가족으로 11 가족을 모집한 경우(88 명)의 검정력이 10배 이상 커지게 된다. 가족의 크기 별로 1) 고전적인 linkage analysis, 2) IBD (identity by descent)를 기반으로 한 penetrance model-free linkage analysis, 3) TDT (transmission disequilibrium test) 분석 등등에 실제 기여하는 표본수 및 표본효율(한 사람 당의 검정력)을 Table 3에 제시하였다.

가족 단위의 코호트는 약 1만 5천 ~ 2만 명 정도의 규모를 목표로 할 경우 Table 2와 동일한 가정 하에서 대부분의 주요 질환들에 대한 분석이 가능할 것으로 추정되며, 약 4천~5천명의 인원을 기존의 쌍둥이-가족 연구에서 [22], 약 3천 명 정도를 영유아 및 소아를 대상으로 하는 성장발달 코호트에서, 또한 나머지를 지역사회 단위 코호트의 가족 하부단위와 일반 코호트에서 가족력이 있는 환례들의 가족쌍의 모집 등의 형태를 통해서 확보해 갈 수 있을 것으로 전망된다.

6. 국가 건강정보를 활용한 추구조사의 전략

기존의 코호트는 추구조사에 대한 명확한 지침과 전략을 가지고 있지 못하다. 악성종양에 대해서는 암등록을 활용할 수 있지만, 다른 질환의 발생을 확인하기 위해서는 일일이 재접촉을 시도해야 한다. 현재는 추구조사에 대한 응답률이 어느 정도인지에 대한 기본 자료도 확보되어 있지 못하다. 예를 들어, 만일 2년 주기로 70%만 추구조사가 된다고 가정하면, 이론상으로는 5년 후에는 약 41%, 10년 후에는 16.8%의 대상만이 추구조사가 되고 있다는 계산이 가능하다. 한 사람 한 사람을 연구에 참여하기 위해서 투자된 자원과 노력을 감안할 때, 추구조사의 방법론을 확보하는 것은 코호트의 성패를 좌우하는 핵심문제라고 생각된다.

우리나라는 1980년대 후반부터 20여 년에 이르는 건강정보가 국가의 주도로 건강보험에 의해서 관리되어 오고 있다. 영국의 UK Biobank가 영국의 국가건강보장제도(National Health Service, NHS)를 통해서 질환의 발생을 거의 완전하게 파악할 수 있다는 자신감을 바탕으로 추진되고 있고 [7], Iceland의 deCODE biobank가 역시 국민들의 건강-질병 정보와 결합되어 엄청난 부가가치를 창출하고 있는 사례 [9]들은 우리나라의 국가 자료원이 활용되어야 할 필요성을 웅변해 주고 있다. 물론, 우리나라의 건강보험자료가 실제 추구조사에 활용되기 위해서는 제도적인 차원의 문제나 개인의 동의를 구하는 윤리적인 차원의 문제뿐만 아니라, 자료를 정확하게 활용할 수 있기 위한 방법론의 개발도 필요하다. 그러나 이러한 과학적인 차원의 문제들은 활용 가능성만 열린다면 큰 문제 없이 해결될 수 있을 만큼 여러 연구자들에 의해서 활용방안이 개발되고 경험이 축적되어 왔다 [23-25].

현재 건강보험자료 등이 개인정보를 보호하는 장치가 부족하기 때문에, 사용을 제한하기 위한 입법조치들이 제안되고 있으나, 개인정보의 철저한 보호와 광범한 공익 활용계획을 수립하는 것은 동전의 양면이다. 즉, 활용을 전제로 할 때 개인정보가 보호될 만큼의 기술력은 이미 확보되어 있지만, 지금처럼 활용이 제한될 때는 개인정보 보호에 대한 필요성도 간과되고 개인정보 유출사고가 발생될 수도 있는 것이다. 유전체 코호트는 물론, 국가의 보건의료 관련 연구-사업에서, 1) 동의를 구한 사람들에 대한 질환발생의 확인, 2) 개인 식별이 불가능하도록 암호화 된 공익목적의 데이터베이스가 구축되는 일(예, 약물부작용의 모니터링 혹은 질병 발생의 위험요인 평가) 등은 시급한 과제이다. 단지 유전체 코호트 연구를 위해서 뿐만이 아니라 세계적으로도 무한한 활용가치를 인정받고 있는 우리나라의 소중한 지식 자료원으로서, 필요한 공익활용을 위해 증지가 모아지고 활용방안이 추구되어야 할 것이다.

결 론

우리나라의 유전체 코호트는 우리가 가

지고 있는 연구 환경의 장단점을 면밀히 분석하여, 국제적인 공동 노력으로 건설되고 있는 "질병정복 기념탑"의 주춧돌이 될 수 있도록 모든 노력이 경주되어야 한다. 예를 들어, 우리나라에는 서구에서는 볼 수 없는 매우 낮은 수준의 콜레스테롤 값(160 mg/dl 혹은 4.13 mmol/l) 을 가진 인구집단이 전체의 15%에 이른다. 서양에서는 전체 인구의 1% 미만이 이러한 수준에 있기 때문에 낮은 콜레스테롤 수준에 있는 사람들의 건강영향에 대한 연구는 원천적으로 불가능하다. 또한, 우리나라는 생활양식의 서구화가 진행되었지만 여전히 전통적인 생활양식이 남아 있어서 생활환경에서 비교적 큰 폭의 다양성이 공존하고 있다. 이러한 특성은 적극적으로 개발되어야 하며, 환경적인 요인뿐만 아니라 한국인의 유전적인 특성에 대해서도 적극적으로 국제적인 안목에서 평가되어야 한다.

유전체 코호트가 다른 나라에 비해서 기존에 확보된 대규모 코호트가 없기 때문에 거의 대부분을 새로 구축해야 하는 것은 전체 연구 예산이 선진국에 비해서 매우 부족한 우리나라에서는 특히 중요한 연구 자원의 독과점으로 비추어질 수 있다. 유전체 코호트가 장기적으로 국가의 유전체, 단백체 및 기타 관련 연구들과 새로 개발되는 과학기술을 직접 지원할 수 있는 강력한 연구자원이 된다는 것을 광범위한 학문 영역의 학자들과 공유할 필요가 있다. 또한, 미래의 잠재적인 사용자들로부터, 향후에 어떠한 연구개발에 활용될 수 있어야 하는지, 이를 위해 무슨 내용들이 지금 조사 · 확보되어야 하는지에 대한 의견수렴과 자문을 적극적으로 받아야 한다. 한편, 다른 나라에 비해서 늦게 출발한 반면 새로운 프로토콜과 더욱 미래지향적인 전략을 가지고 코호트가 진행될 수 있다는 점은 우리가 극대화해야 할 장점이라고 할 수 있다.

지금까지, 유전체 코호트의 목표 규모, 대상질환, 측정내용과 표준화, 생체자원의 체계적인 보관과 정도관리, 효율적인 성과를 얻기 위한 가족 단위와 특수집단의 포괄, 그리고 국가정보원을 적극적으로 활용한 추구조사 체계를 만들어야할 시급성에 대해서 언급하였다. 또한, 우리나라의 코호트가 국제적인 기여를 하기 위한 정밀한 검토의 필요성, 그리고 학계에서의 광범한 의사소통의 시급성을 지적하였다. 이러한 현안의 해결은 현재 구축되고 있는 유전체 코호트가 우리나라 국민들의 건강증진을 위해서, 나아가 세계의 과학자들과 함께 인류의 질병정복을 위해서 실제로 중요한 역할을 담당하는 연구자원이 될 수 있기 위한 필요조건이 될 것이다.

참고문헌

1. Collins. FS. The case for a US prospective cohort study of genes and environment. *Nature* 2004; 429(6990): 475-477
2. Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. *Science* 2002; 298(5602): 2345-2349
3. Terwilliger JD, Weiss KM. Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann Med* 2003; 35(7): 532-544
4. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. *Nature* 2004; 429(6990): 446-452
5. Potter JD. Toward the last cohort. *Cancer Epidemiol Biomarkers Prev* 2004; 13(6): 895-897
6. Bentley DR. Genomes for medicine. *Nature* 2004; 429(6990): 440-445
7. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. *Pharmacogenomics* 2005; 6(6): 639-646
8. Triendl R. Japan launches controversial Biobank project. *Nat Med* 2003; 9(8): 982
9. Hakonarson H, Gulcher JR, Stefansson K. deCODE genetics, Inc. *Pharmacogenomics* 2003; 4(2):209-215
10. Sung JH. Years of life lost and health priority in Korea. *Korean J Epidemiol* 1997; 19(2): 200-209 (Korean)
11. Murray CJ, Lopez AD, Jamison DT. The global burden of disease in 1990: summary results, sensitivity analysis and future directions. *Bull World Health Organ* 1994; 72(3): 495-509
12. 건강보험심사평가원. 우리나라의 보험 청구자료의 진단명 정확도에 대한 연구 보고서. 건강보험심사평가원; 2003. p. 10-21
13. Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet* 2006; 7(10): 812-820
14. Yoo KY, Shin HR, Chang SH, Choi BY, Hong YC, Kim DH, Kang D, Cho NH, Shin C, and Jin YW for the Korean Genome Epidemiology Society. Genomic epidemiology cohorts in Korea: Present and the future. *Asian Pac J Cancer Prev* 2005; 6(3): 238-243
15. Willett WC, Blot WJ, Colditz GA, Folsom AR, Henderson BE, Stampfer MJ. Merging and emerging cohorts: not worth the wait. *Nature* 2007; 445(7125): 257-258
16. Collins FS, Manolio TA. Merging and emerging cohorts: necessary but not sufficient. *Nature* 2007; 445(7125): 259
17. Jun JK, Gwack J, Park SK, Choi YH, Kim Y, Shin A, Chang SH, Shin HR, Yoo KY. Fasting serum glucose level and gastric cancer risk in a nested case-control study. *J Prev Med Pub Health* 2006; 39(6): 493-498 (Korean)
18. Peltonen L. GenomEUtwin: A strategy to identify genetic influences on health and disease. *Twin Res* 2003; 6(5): 354-360
19. Ronningen KS, Paltiel L, Meltzer HM, Nordhagen R, Lie KK, Hovengen R, Haugen M, Nystad W, Magnus P, Hoppin JA. The biobank of the Norwegian mother and child cohort Study: A resource for the next 100 years. *Eur J Epidemiol* 2006; 21(8): 619-625
20. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998; 62(5): 1198-1211
21. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; 30(1):97-101
22. Sung J, Cho SI, Lee K, Ha M, Choi EY, Choi JS, Kim H, Kim J, Hong KS, Kim Y, Yoo KY, Park C, Song YM. Healthy twin: A twin-family study of Korea - Protocols and current status. *Twin Res Hum Genet* 2006; 9(6): 844-848
23. Jee SH, Sull JW, Park J, Lee SY, Ohrr H, Guallar E, Samet JM. Body-mass index and mortality in Korean men and women. *N Engl J Med* 2006; 355(8): 779-787
24. Ebrahim S, Sung J, Song YM, Ferrer RL, Lawlor DA, Davey Smith G. Serum cholesterol, haemorrhagic stroke, ischaemic stroke, and myocardial infarction: Korean national health system prospective cohort study. *BMJ.* 2006; 333(7557): 22
25. Kim HJ, Lee SM, Choi NK, Kim SH, Song HJ, Cho YK, Park BJ. Smoking and colorectal cancer risk in the Korean elderly. *J Prev Med Public Health* 2006; 39(2): 123-129 (Korean)