유전자 네트워크 기반 전장유전체 연구의 새로운 시도

New approach for Genome-Wide Association Study based on genetic networks

질병관리본부 국립보건연구원 유전체센터 바이오과학정보과

Ⅰ. 들어가는 말

최근 고집적(high throughput) 단일염기다형성(SNP) 실험기술의 발전으로 분석 비용이 감소하면서 대규모 Genome-Wide Association Study(GWAS)가 실현되었다. GWAS 분석을 이용하면 대규모 인구집단의 유전체 데이터에서 특정 질병과 연관된 유전적 변이를 신속하게 탐색할 수 있다. 이러한 GWAS 분석들은 현재 천식, 암, 당뇨 등 복합 질병들의 유전적 변이를 찾는데 주로 이용되고 있고, Wellcome Trust Case Control Consortium(WTCCC), SNP Health Association Resource(SHARe) 컨소시엄 등에서 2007년을 기점으로 폭발적인 결과들을 쏟아내고 있다[1]. 대표적인 예라고 할 수 있는 WTCCC 연구에서는 유럽의 정상인 3,000명과 7개의 질환별로 각각 2,000명의 환자군으로 디자인되어 역학조사가 이루어졌으며, 50만 개의 SNP 유전체에 대한 기반 분석이 진행되었다[2]. 질병관리본부 유전체센터에서도 한국인 1만 명에 대한 50만 개의 SNP 유전체 정보를 생산하고, Korea Association Resource(KARE) 컨소시엄 구성을 통해 국내 연구자들이 이 자료를 활용하여 많은 연구 성과를 내고 있다.

이처럼 생명과학 분야에서는 21세기에 들어와 실제 유전정보를 바탕으로 한 유전체 분석과 그 외 역학정보 등과의 통합분석을 통하여 질병의 실제 원인을 밝혀나가고 있다. 1997년부터 2009년까지 투고된 논문을 토대로 질병 연구와 유전자 연구들의 증가 경향을 보면 유전자와 질병의 연구는 계속해서 활발하게 이루어지고 있는 것을 알 수 있다(Figure 1).

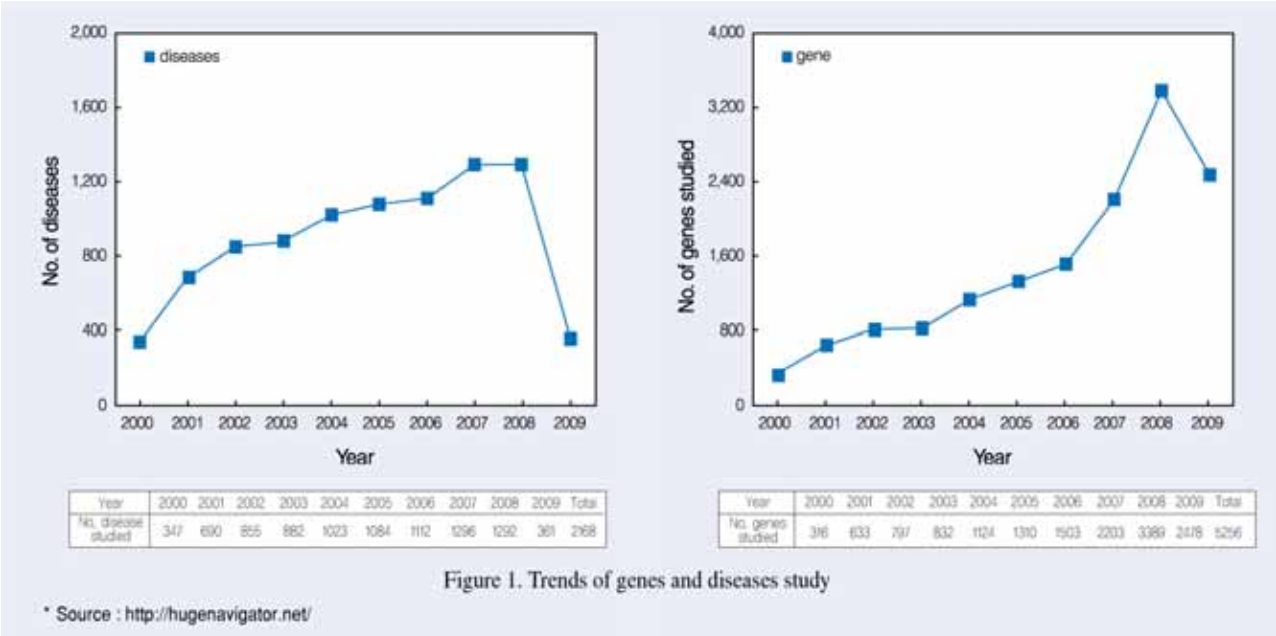

Figure 1. Trends of genes and diseases study

* Source : http://hugenavigator.net/

대부분의 생물학 연구는 단기간에 실용화되어 실생활에 적용되기 어렵지만 이와 같은 질환 지표(marker) 발굴 관련 연구들은 생물학적 기전을 밝히는 많은 연구단계를 건너 뛰어 실용화, 산업화가 빠르게 이루어지고 있다. 대표적인 예로 유전자와 질병연구를 통해 얻어진 질병과 건강에 관련된 유전자 정보는 선진국을 중심으로 곧바로 인터넷을 활용한 사업모델 Direct to Consumer(DTC)로도 전환되고 있다[3].

미국의 대표적인 DTC 유전자검사 업체인 23andMe의 경우는 세계적인 인터넷 업체인 구글의 투자를 받아 큰 화제가 되었다. 이 업체는 최근 유전자검사 비용을 999달러에서 399달러로 대폭 낮추는 등 공격적인 전략에 나섰으며, 2008년 타임지에 의해 '올해의 발명품'으로 선정되기도 하면서 이 분야의 과학적, 산업적인 변화의 흐름을 보여주고 있다. 이외에도 Navigenics 등 여러 기업들이 다양한 유전체 정보 기반 사업모델을 전개하고 있으며, 현재의 SNP 칩 중심 모델에서 향후 유전체 서열 중심으로 변화될 예정이다(Table 1).

Table 1. Example of DTC genetic testing (2009. 4)

Company	Nation	Website	Genetic testing	Testing cost(USD)
Navigenics	U.S.A.	www.navigenics.com/	Whole Genome Analysis	2,499 +499 (year)
23andMe	U.S.A.	www.23andme.com/	SNP Genetic Analysis (disease)	399
deCode Genetics	U.S.A., Island	www.decodeme.com/	Genetic risk for 38 diseases	985, 225, 195
Nico Test	United Kingdom	www.nicotest.com/	Nicotine Metabolism	£150
Genetic Health	United Kingdom	www.genetic-health.co.uk/	Obesity, Diabetes and	£180
			Weight Loss	£293
			Nutrition gene	£295
			Pharmaco gene	£825

II. 몸 말

질병관리본부 국립보건연구원 유전체센터에서는 다양한 코호트 기반사업을 통해 당뇨를 비롯한 몇몇 만성질환에 대한 임상, 역학 정보와 유전체 변이 정보를 보유하고 있으며 많은 유전체 연구자들이 질환 관련 유전자 연관성을 찾기 위해 심혈을 기울이고 있다. 그 결과 유전체 센터 내에서도 KARE 컨소시엄을 통해 비만, 키, 혈당, 골밀도, 혈압 및 질병과의 관계에 대한 결과를 발표한 바 있으며[4], 유럽의 WTCCC 컨소시엄에서도 7개의 질병에 해당하는 연관 유전자를 찾아냈다(Figure 2).

기존 유전자 연구의 기술적, 비용적인 측면의 문제점이 하나씩 해결되면서 많은 연구자들의 관심도 급증하고 있으며 Figure 2에서 보는 바와 같이 질병 연관성에 대한 의문이 상당부분 해소되어가고 있다. 그러나 생산된 데이터의 양에 비해서 유의한 연관성을 가지는 SNP의 수는 생각보다 그리 많지 않다. 그 원인 중 하나는 현재 대부분의 연구에서 환경 요인이나 유전적 인자들의 복합적인 작용을 고려하지 않고 독립적인 SNP 각각에 대해 유전통계학적 방법을 적용하여 질병 감수성을 평가하고 있기 때문이다[5].

이를 해결하기 위한 방법으로 우리는 몇 가지 생물정보학적 분석방법을 적용하여 유전자간의 상호작용 및 세포 내 기능을 고려한 질환 지표 탐색 연구를 하고자 하였다. 인간 전장 유전체 전체를 대상으로 계산하기 전에 바이오과학정보과는 작은 데이터 셋을 이용하여 선행 분석을 하였다. 세계보건기구(World Health Organization)의 기준을 따라 공복혈당 126 ㎖/dℓ 이상인 당뇨 환자 462명과 정상인 456명에서 선정된 87개의 유전자에 존재하는 487개의 SNP 중에서 데이터 전처리를 통해 408개 SNP를 데이터 셋으로 정하였다. 복수의 SNP들이 조합을 이루어 보여주는 질환에 대한 감수성을 찾기 위해 다양한 SNP 조합의 판별분석능력을 Support Vector Machine(SVM)으로 평가하였다[6]. 408개 SNP 전체를 사용하는 경우, 환자와 정상인을 판별 분석하는 능력은 대부분의 기존 알고리즘에서 50% 전후의 결과밖에 성능을 내지 못했지만, 408개중 14개의 특정 SNP들로 이루어진 최적 조합을 찾은 결과 65.3%까지 정확도가 향상되었다. 900여명의 집단을 남성, 여성의 소집단으로 나누어서 마찬가지 방법으로 분석한 결과, 각 집단별로 특유의 SNP 조합이 존재한다는 것도 알 수 있었다(Table 2).

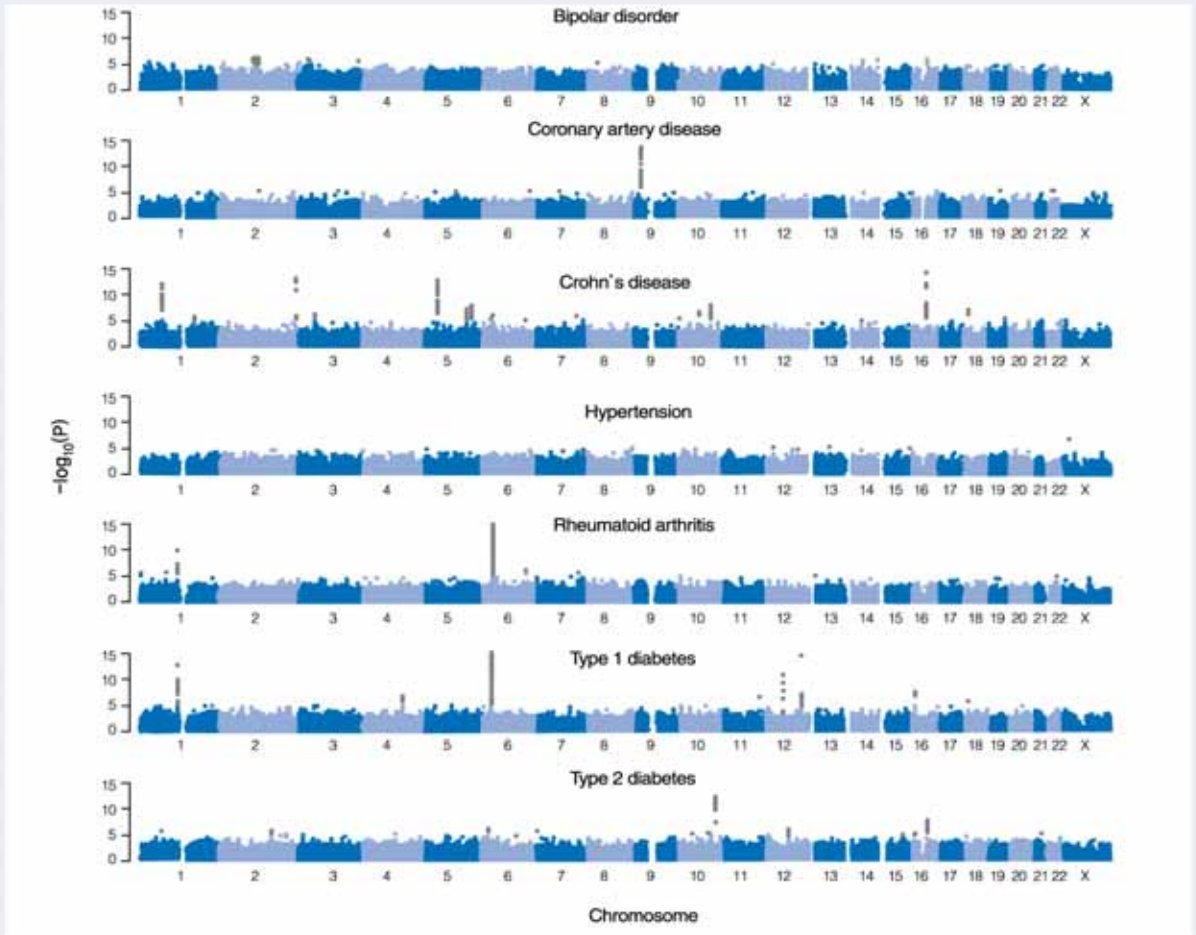

Figure 2. Genome-wide scan for associations of SNPs with each of the seven diseases

* Source : http://www.wtccc.org.uk/info/070606.shtml

Table 2. Prediction rates of the SVM classifiers with different target populations

Target population	Sensitivity	Specificity	Overall accuracy	No. of SNPs for each combination
Total	0.567	0.739	0.653	14 SNPs
Men	0.714	0.704	0.709	10 SNPs
Women	0.715	0.696	0.706	19 SNPs

14개의 SNP 조합이 속하는 유전자들의 Protein-Protein Interaction(PPI)을 조사한 결과(GENOME NETWORK PLATFORM 데이터베이스 이용하여 결과를 얻음), IRS1, PPARD 단백질 등은 4개 이상의 단백질들과 상호작용을 하는 일종의 허브 역할을 할 가능성이 높은 것으로 보이며, 이러한 조합 내의 단백질들을 연결시켜 주는 다른 단백질들인 JAK1, SRC, IL4R 등이 중요한 역할을 가짐을 알 수 있었다(Figure 3-a). 각 서브그룹은 당뇨병 환자들에 있어 성별에 따라 특이적인 유전자 작용이 존재할 것이라는 가정에서 분석되었으며, 최적 조합 SNP들의 패턴이 다르게 나타남을 알 수 있었다. 네트워크를 보면 막에서 인슐린 인식을 도와주고 전달해주는 유전자 종류가 세 개의 그룹에서 중복된 결과를 보이며 그 이외의 유전자는 그룹별로 다르게 나타났다(Figure 3-b, c). 이는 당뇨병 발병에 핵심이 되는 유전자 군과 성별에 따라 다르게 작용될 가능성이 있는 유전자 군의 후보들을 밝혀내는 방법으로 적용할 수 있을 것이다.

유전체센터에 축적된 유전변이정보를 이용하여 모든 조합을 탐색하는 것은 현실적으로 불가능할 정도의 계산량이 요구되므로 유전체센터에서는 현재 보유하고 있는 슈퍼컴퓨터 리소스를 활용하여 초고속 탐색이 가능한 병렬처리 프로그램을 SVM 기계학습이론에 적용하는 방안을 모색 중이다. 이러한 방안은 새로운 GWAS 기법의 하나로서 대용량 전장 유전체 정보를 대상으로 기계학습이론을 적용한 질환 지표 발굴이 유전자 상호작용의 영향성을 확인하는 방향으로 전개될 것이다.

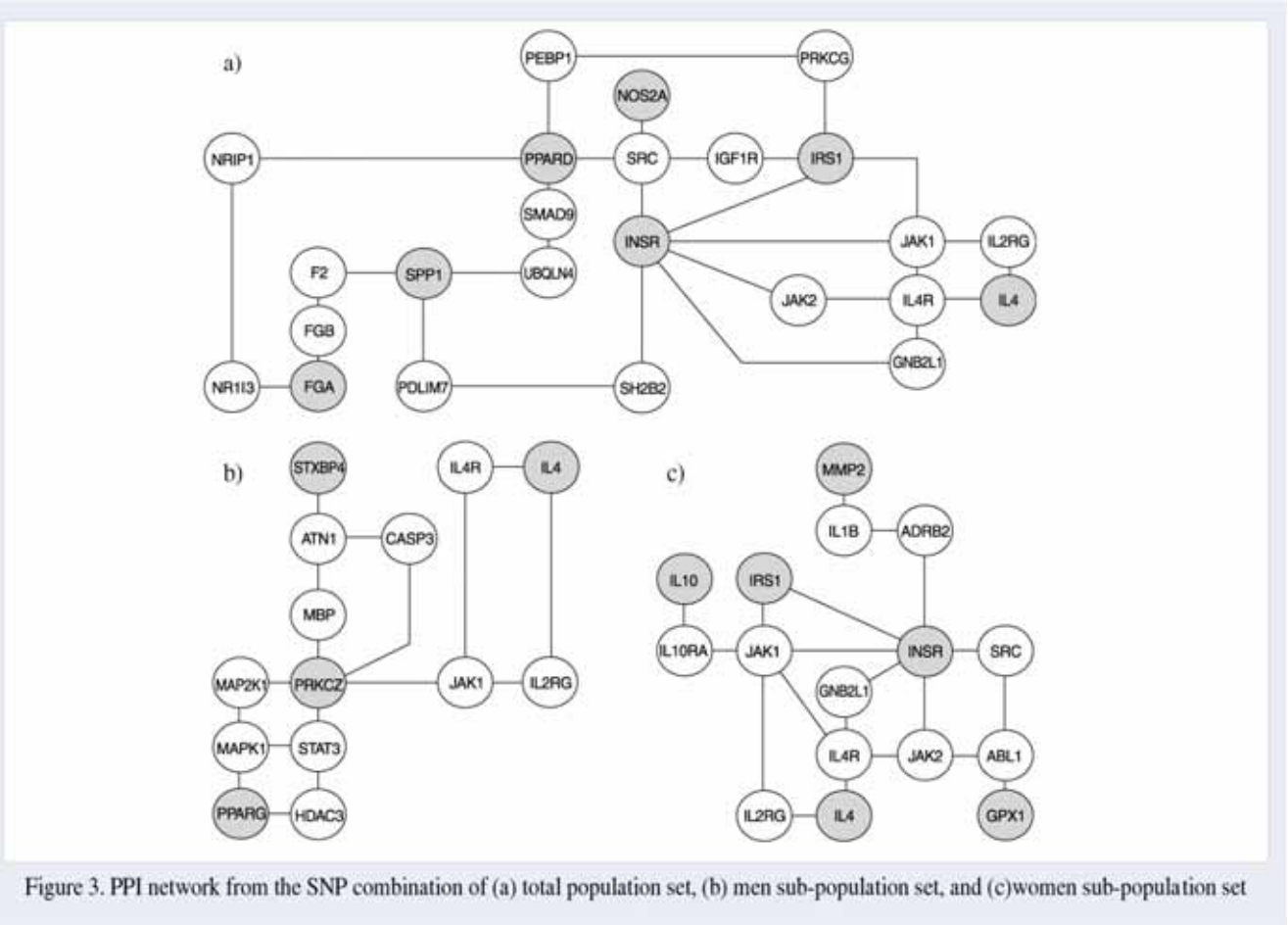

Figure 3. PPI network from the SNP combination of (a) total population set, (b) men sub-population set, and (c)women sub-population set

III. 맺는 말

이와 같은 연구는 생물학적 지식기반과 유전자 상호작용 등을 고려하여 질환 지표를 발굴하는 기법을 전장유전체의 대용량 SNP 데이터에 대하여 최초로 적용하게 된다. 유전체 정보와 질병의 연관성이 계속해서 밝혀지고 그 결과가 직접적으로 산업화와 연결되는 최근의 세계적인 추세를 고려할 때, 국내에서도 이러한 연구가 대규모로 신속하게 수행될 필요가 있다고 여겨진다.

복합질환의 특성인 유전자의 복잡한 상호작용이 실질적으로 반영된 질환 지표 탐색이라는 관점에서 질병이나 임상형질에 영향을 끼치는 유전변이 기전 해명에 기여할 수 있으며, 질병과 연관되는 유전변이 정보에 대한 이러한 분석은 곧 다가올 미래의 개인 유전체시대를 대비하는 기초가 될 것이다. 이것은 인간유전체 정보와 질병을 연결하는 매우 중요한 연결고리를 제공하며, 복합질환의 광범위한 위험요인 탐색을 현실화시켜서 각종 질환 예방과 치료에 필요한 유전적, 환경적 요인에 대한 궁극적인 통합분석으로 이어지기를 기대한다.

IV. 참고문헌

1. A Catalog of Published Genome-wide Association Studies, http://www.genome.gov/26525384
2. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447;661-78.
3. http://www.genome.gov/24519851
4. Yoon Shin Cho. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nature Genetics.2009;41; 527-534.

5. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 6;456(7218);18-21
6. Joachims,T. Making large-scale SVM learning practical. Advances in kernel methods-support vector learning, Sch쉢opf,B., and Burges,C., and Smola,A. (eds), MIT Press 1999.