예방의학회지 제40권 제2호(2007년 3월)
J Prev Med Public Health 2007;40(2):102-107

특 집

유전체 코호트 연구를 위한 대용량 염기서열 분석

박웅양

서울대학교 의과대학 생화학교실 및 인간유전체연구소

High Throughput Genotyping for Genomic Cohort Study

Woong-Yang Park

Department of Biochemistry and Human Genome Research Institute, Seoul National University College of Medicine

Human Genome Project (HGP) could unveil the secrets of human being by a long script of genetic codes, which enabled us to get access to mine the cause of diseases more efficiently. Two wheels for HGP, bioinformatics and high throughput technology are essential techniques for the genomic medicine. While microarray platforms are still evolving, we can screen more than 500,000 genotypes at once. Even we can sequence the whole genome of an organism within a day. Because the future medicne will focus on the genetic susceptibility of individuals, we need to find genetic variations of each person by efficient genotyping methods.

J Prev Med Public Health 2007;40(2):102-107

Key words : Genotype, Sequence analysis, DNA, Microarray analysis, Single nucleotide polymorphism

서 론

2001년 인간게놈프로젝트의 완성과 함께 우리는 인간을 비롯한 수 십 여종의 생명체에 대한 전체 유전체 수준에서 유전정보를 확보하게 되었다. 유전체에 대한 정보를 확보할 수 있게 됨과 동시에 대규모의 염기서열을 분석할 수 있는 기술들이 발달되고 있다. 이와 같은 대용량 유전체 염기서열 분석기술은 인구집단의 유전적 특성분석이 빠르고 정확하게 될 수 있도록 하였으며, 세계적으로 대용량 유전체 분석을 기반으로 한 대규모 코호트 연구들이 진행되고 있다. 아시아인종은 유전적으로 백인 및 흑인종에 비해 고유하며 한국인에 대한 유전체 코호트 연구는 우리의 문제 뿐 만이 아니라 아시아인종 전체를 위한 작업이라고 할 수 있다.

1. 유전자 염기서열 분석

인간의 다양성을 설명할 수 있는 방법은 인구집단 또는 개인에게 주어진 서로 다른 환경의 영향과 함께 3-4만 년 전부터 시작된 인간의 유전자 염기서열 다형성에 기인한다고 할 수 있다. 즉, 인간의 행동과 생리를 설명하는 환경적 요인과 유전적 요인으로 나누어 볼 수 있다. 몇 가지 지표만으로는 분석이 어려운 복잡한 환경적인 요인과 달리 인간의 유전체내에 국한된 유전적 요인은 4가지 염기에 의해 결정되기 때문에 "분석"가능한 영역내에 있다. 따라서 유전적 요인이 차지하는 인구집단 또는 개인간의 생리현상의 차이 또는 질병 패턴의 원인에 대해서는 이들 염기서열의 분석에 의해 가능하다고 할 수 있다.

인간 질병에 대한 유전적 요인을 이해하기 위한 염기서열 분석은 2001년 발표된 인간게놈프로젝트의 완성이후로 폭발적으로 증가되었다. 이전에는 특정한 질병을 가진 가계분석을 통해 유전자를 클로닝하고, 질병과 관련된 돌연변이를 찾는 작업은 소규모로 진행되어 왔다. 이러한 방법으로 1,500여개의 질환 유전자를 찾을 수 있었으며, 이들에 대한 자료를 OMIM과 같은 DB에 정리되어 있으며, 주로 monogenic disease에 대한 정보를 얻을 수 있다. 하지만 이러한 방법으로 유전자를 찾기 위해서는 유전자지도와 landmark가 될 수 있는 genetic marker가 중요하다. 2001년 인간게놈 프로젝트는 질환 유전자 클로닝을 위해 두 가지 중요한 수단을 제공하였다. 첫 번째는 완벽한 유전자지도를 제공함으로써 유전자 또는 마커를 손쉽게 찾을 수 있도록 한 것이다. 두 번째로는 대용량 유전자 염기서열 분석 기술을 가능하게 한 것이다. 인간 유전체 전체 염기서열을 물리적으로 분석하기 위한 대용량 분석법의 개발과 함께, 이러한 정보를 분석할 수 있는 DNA chip과 같은 초고속 분석법이 가능하게 되었다.

인간의 다양성은 질환 감수성의 차이를 설명할 수 있다. 즉 개인이 서로 다른 것처럼 질환에 대한 감수성이 다른 것이다. 이는 앞서 얘기한 바와 같이 유전적 요인에 의해 설명할 수 있으며, 유전적 요인은 개인간의 유전자 염기서열의 차이들의 집합으로 설명할 수 있다. 따라서 인구집단과 개인의 염기서열을 분석하면 각 개인의

이 연구는 질병관리본부 학술용역사업(2006-347-2400-2440-215)으로 수행된 내용에 근거한 것임.
책임저자 : 박웅양(서울시 종로구 연건동 28번지, 전화 : 02-740-8241, 팩스 : 02-744-4534, E-mail : wypark@snu.ac.kr)

질환 감수성을 찾을 수 있을 것이다. 이러한 노력은 잘 알려진 바와 같이 deCode Genetics사의 전략과 일치하는 것이다. 즉, 특정한 인구집단, 특히 아이슬란드와 같이 최근 수백 년 간 고립된 지역에서 유전적으로 고유한 특질을 지닌 인구집단의 염기서열을 분석하고 이들 집단에서 질병 패턴을 분석함으로써 유전적 질환 감수성을 파악할 수 있다는 것이다. 실제로 이들의 노력은 50여개의 질환유전자와 이에 대한 돌연변이, 유전자 다형성을 찾아 보고한 것으로 결실을 맺고 있다.

유전적 요인을 분석하기 위해서는 염기서열을 정확하고 빠르게 분석할 수 있어야 한다. 특히 유전체코호트 사업과 같이 수 천 명의 전체 유전체 염기서열을 정확히 분석하기 위해서는 Table 1에서 제시한 효과적인 염기서열 분석법이 필요하다. 본 논문에서는 현재까지 보고된 각종 genotyping methods를 분석하고, 효율적인 대용량 고효율 분석법에 대하여 설명하고자 한다.

Table 1. Ideal method for high throughput genotyping

이상적인 유전자형 분석법은
1. 유전자 염기서열로부터 쉽게 개발될 수 있고,
2. 전문가가 염기서열 특이적으로 저렴하게 빠르게 분석법을 최적화할 수 있으며,
3. 적은 양의 DNA 시료로부터 대량의 분석이 가능하고,
4. 완전 자동화내지는 최소한의 수작업에 의해 분석가능 해야 한다.
5. 간단하고 자동화된 정확한 염기서열 해독이 가능해야 하며,
6. 수백 개로부터 수백 만 개의 유전자형 분석으로 쉽게 확장할 수 있고,
7. 최적화된 유전자형 분석에 대해 기기 및 시약에 드는 비용이 적어야 한다.

본 론

1. SNP 분석기술 현황

개인간의 유전자 다형성중의 일부는 SNP으로 설명할 수 있다. Size variation과 같은 다른 형태의 유전자 다형성에 비해 고르게 분포하며, 밀도가 높아 유전자 마커로서의 유용성이 높다. 하지만 2개 또는 4개의 유전자형으로 제한되어 다른 유전자 마커에 비해 다양성이 부족하기 때문에 많은 수의 SNP 염기서열 분석이 필요하다. SNP의 장점은 염기서열 분석을 공정화하여 대용량 초고속 분석이 가능하다는 것이다. 즉, 특정한 염기서열을 대량으로 분석할 수 있는 방법을 개발함으로써 동시에 SNP을 대량으로 분석할 수 있게 된다면, 쉽게 각 개인의 유전적 요인을 분석할 수 있을 것이다.

현재 보고된 SNP은 3천만 개에 이르며, 이중에도 인구집단에서 5%이상의 minor allele frequency를 갖는 SNP의 수도 7백만 개에 다다른다 [1]. 물론 이는 코카시안 인구집단을 대상으로 한 것이며, 아시아 인종이나 아프리카 인종의 경우 다른 조합의 SNP 또는 새로운 SNP이 있을 것으로 예상할 수 있다. 현재 분석 가능한 대용량 분석법으로는 50만개의 염기서열을 동시에 분석할 수 있는 것으로 이는 전체 SNP의 10%이내에 해당하는 숫자이다. 이들 SNP을 통하여 질환군과 정상 대조군간의 연관성 분석 (association study) 으로 질환 감수성 유전자를 찾을 수 있다. 또한 질환 유전자 클로닝을 위하여 유전자자리를 찾는 작업에서 최종적으로 세밀하게 찾기 위해서는 조밀한 유전자 마커가 필요한데, 이 때 SNP은 매우 유용하게 사용될 수 있다. 즉, 기존의 STR marker 와 함께 사용함으로써 유전자자리를 효율적으로 찾을 수 있는 방법을 제공하고 있다. 최근에는 인체 질환에서 copy nmumber variation과의 관련성을 위해 SNP genotyping에서의 연구결과가 array CGH등과 연계되어 매우 중요하게 보고되고 있다 [2].

2. 분석의 범위에 따른 genotyping 법의 선택

각각의 SNP genotyping 방법에 따라 분석 가능한 SNP 의 수와 효율이 다르다. 가장 먼저 본인이 분석하고자 하는 SNP의 개수와 함께, 시료의 숫자를 정하고 이로부터 가장 적합한 분석법을 찾아야 한다. 즉, 10개 이하의 SNP에 대해 수천 개의 샘플을 분석하기 위해서는 Applied Biosystems사의 TaqMan real-time PCR 방법이 제일 유용하다. 이외에 Sequenom사의 MassARRAY 시스템도 1,000여개 이내의 샘플에서 SNP을 분석할 때에는 효율적이다. 또한 기존에 사용되는 각종 genotyping 방법들, 즉 pyrosequencing이나 sequencing-by-synthesis에 의존하는 방법들도 100개 이내의 샘플을 위해서는 모두 유용하다고 할 수 있다. MegAllele과 같은 방법은 10,000개의 SNP을 분석할 수 있으며, Affymetrix사의 경우 50만개의 SNP을 동시에 분석할 수 있는 GeneChip Human Mapping 500K Array Set를 제공하고 있다. 이와 같은 방법들은 전체 유전체 수준에서 SNP 분석을 가능하게 하는데, 최근에는 Illumina사에서 개발한 BeadArray technology를 이용하여 많은 기관에서 사용하고 있는 실정이다. Table 2에서 정리한 바와 같이 각각의 genotyping 방법들의 특징과 장단점을 이해하고 시료와 대상 SNP의 숫자를 고려하는 것이 매우 중요하다.

여러 genotyping 분석법의 효율성은 자동화 수준과 작업자의 숙련도에 의해 결정된다. Affymetrix GeneChip Scanner 3000 System은 이론적으로 48개의 array로 48개의 시료를 하루에 처리할 수 있다. 이로써 50만개 SNP에 대해 48개의 시료를 하루에 분석하여 2천4백만개의 genotyping이 가능하다. 하지만 데이터 분석을 포함하면 2-3일이 더 소요되며, 실제 시료를 전처리하는 기간까지 포함하면 한 사람이 처리할 수 있는 genotyping은 더 적게 된다. 반면에 Illumina사의 "DNA-to-data cycle"에 의하면 시스템과 데이터 분석의 자동화를 통하여 한 사람이 288 개의 시료까지 처리할 수 있다.

Table 2. Characteristics and throughput for different platforms for genotyping

Platforms	Example(s)	Characteristics	Throughput
Microtiter Plates	TaqMan	Good for a few markers PCR prior to genotyping	Low
Size Analysis by Electrophoresis	SNPlex	Intermediate multiplexing Reduced costs	Medium
Arrays	Illumina Affymetrix	Genotyping directly on genomic DNA Highly multiplexed Reducced costs	High

ABI사의 7900 HT는 384개의 TaqMan assays를 30분 만에 처리하거나 48개의 384 well plate, 즉 18,432개 샘플을 하루에 처리할 수 있다. 반면에 Sequenom사의 경우 high-capacity Autoflux mass spectrometer를 이용하면, 20개의 384-well chip (7,680samples)들을 분석할 수 있다. Pyrosequ- encing을 사용하면 96개 시료를 10분 만에 처리하는데 이를 위해 PCR을 해야 하는 번거로움이 있다.

3. 기존의 Database를 이용한 분석

물론 작은 규모의 분석을 위해 restriction fragment length polymorphism도 사용할 수 있다. 하지만 dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) 을 비롯하여 HapMap (http://www.hapmap.org/) 등의 자료를 이용하여 원하는 질환관련 유전자 또는 감수성 유전자, biomarker 후보를 쉽게 선정하고 이를 대용량 검색을 통하여 한국인 인구집단에서 SNP을 찾는 것은 매우 좋은 전략이 될 것이다.

최근에 발표된 HapMap 자료를 이용한 유전자 다형성 분석으로는 펜실베니아 주립의대의 Cheng 박사는 2005년 Science에 발표한 논문을 들 수 있는데, 이들은 피부색의 다양성에 대한 유전학적 연구를 위해 HapMap 연구 결과를 사용하였다 [3]. 이들은 Zebra fish에서 유전지표로 사용하는 *golden* 유전형에 대한 유전자를 morpholino를 사용하여 클로닝하여 SLC24A5라는 유전자임을 밝혔다. 이들은 SLC24A5의 human ortholog를 찾은 후에 HapMap에서 분석한 다양한 인구집단간의 차이를 찾고자 노력하여 11번째 아미노산이 아프리카 및 아시아인종과 같이 피부가 진한 경우 alanine이고, 유럽인종의 경우 threonine임을 알게 되었다.

이들의 전략은 두 가지로 요약할 수 있는데, 첫 번째는 HapMap과 같은 유용한 DB를 사용하여 관련 genotype을 빠르게 찾을 수 있었다는 것이고, 두 번째는 Zebra fish와 같이 간편한 모델동물을 이용하였다는 것이다. 모든 질환이나 생리현상에 적합한 모델동물을 모두 갖는 것이 쉬운 일은 아니지만, 연구자가 원하는 현상에 대한 적절한 모델동물을 활용하는 것은 우수한 전략 중에 하나라고 할 수 있다.

모델동물과 함께 적절한 인구집단을 사용하는 것도 생각할 수 있으며, 이는 가계분석이나 고립부족을 통한 연구에서 가능할 것이다. 이는 동물모델보다 더 우수한 샘플이 될 수 있으며, 바로 인간의 질환에 적용할 수 있다. 앞서 얘기한 deCode Genetics사의 전략으로, 최근 인구집단에 대한 유전자형 분석으로 STR marker 분석으로 질환유전자를 찾고자 할 때 fine mapping의 방법으로 SNP genotype을 사용하고 있다. 또한 가계분석외에 연관성분석을 위해서 SNP genotyping과 STR 분석을 병행하여 각각의 장단점을 적절히 활용하고 있다. 따라서 유전체 코호트 분석의 경우에도 SNP genotyping외에 다른 marker들의 분석 방법이나 샘플들에 대한 인자를 같이 고려해야 할 것으로 판단된다.

4. 모델동물의 활용

모델동물 자체에서 질환관련 SNP을 분석하는 경우에는 자체의 염기서열에 대한 정보가 필요하다. 최근 염기서열 분석의 결과로 각종 모델동물들의 전체 유전체 염기서열이 밝혀지고 있으나, SNP에 대한 자료는 인간에 비하여 보잘 것 없다. 즉, 생쥐의 경우 약 60만개의 SNP 이 보고되어 있고, 꼬마선충의 경우 천여 개, 소의 경우 2만 여개가, 그리고 쌀의 경우 백칠십만 개의 SNP이 dbSNP에 보고되어 있다. 최근 염기서열 분석이 증가하고 있는 이들 모델동물이나 작물의 경우 더 많은 SNP이 앞으로 보고될 것이고, 이를 유용하게 이용하는 것이 필요하다.

5. Genotyping 비용

최근의 genotyping 분석은 대부분의 대용량분석방법들로 고가의 기기에 대한 투자가 필요하다. 하지만 전반적으로 개별 유전자형 분석을 위한 비용은 비교적 저렴하다. 비용의 산정을 위해 고려해야 할 것으로는 각 SNP에 대한 비용이 아니라 sample에 대한 비용이 얼마나 드는가에 대한 것이다. 즉, 필요에 의해 sample의 수를 늘려야 하는 경우 추가 비용을 계산하여야 하는 경우가 많다. 또한 높은 재현성을 가진 방법은 통계처리를 위한 실험 반복의 횟수를 줄여서 전체 비용을 줄일 수 있다.

최근 미국 버지니아주의 Bioinformatics사에서 분석한 결과에 의하면 보고된 SNP 전체의 40%가량은 외부용역 또는 기관의 중앙지원시설을 통해 분석하였다고 한다. 이는 비용을 줄이기 위한 방법이라고 해석할 수 있다. 예를 들어 영국 Oxford에 있는 Wellcome Trst Center에서는 Illumina사의 beadArray를 유전체 검색법으로 사용하고, Sequenom사의 MassARRAY를 fine mapping의 방법으로 사용한다. 각각 3-4억 원의 기초투자가 필요한 장비들이다. 미국 보스턴의 MIT대학내 Broad Institute에서는 Sequenom과 Affymetrix platform 을 사용한다. Affymetrix에 비해 Illumina의 array는 자신이 원하는 SNP을 선택할 수 있기 때문에 정해진 SNP에 집착하기보다 자신의 SNP list를 가진 경우 유리할 수 있다.

Table 3. Cost for genotyping based on SNP number

SNP No.	Cost ($/genotype)	Cost per 1,000 sample ($)
5~ 10	0.60	6,000
48~ 96	0.25~0.30	~ 29,000
384~ 1,536	0.08~0.15	57,600~122,880
300,000~500,000 (defined format)	0.002	400,000~800,000
10,000 ~20,000 (custom formats)	0.03	>250,000

6. 각종 대용량 genotyping 법의 비교

1) 대용량 염기서열 분석

2003년 9월 Venter 박사는 J. Craig Venter Science Foundation을 통해 인간유전체 염기서열 분석을 천불에 가능하게 할 수 있는 기술을 개발하는 사람에게 50만 불을 수여하겠다고 하였다. 이어서 미국 California주 산타모니카에 있는 X Prize Foundation에서는 5백만 불 내지 2천만 불의 상금을 걸었다. 2004년에는 미국 NIH의 국립유전체연구소에서 Collins박사는 포유동물 크기의 유전체를 초기에는 10만 불, 궁극적으로는 천불에 분석할 수 있는 기술 개발을 위해 7천만 불, 즉 700억 원의 연구비를 투자하겠다고 공표하였다. 이러

한 노력에 힘입어 대용량 초고속 분석법들이 소개되고 있다.

대표적으로 454 Life Science사에서는 bead sequencing법으로 짧은 크기의 염기서열을 빠르게 pyrosequencing으로 분석하는 시스템을 구축하였다 (Figure 1). 이들은 *Mycoplasma genitalium*의 2천5백만 개 염기를 99.4%의 정확도를 가지고 분석하여 발표하였으며, 2.1Mb 유전체를 가진 *Mycobacterium tuberculosis*를 분석하여 2005년 Science지에 발표하였다 [4]. Pyrosequencing의 문제로 100bp정도의 짧은 염기서열을 분석하지만 24시간 내에 수십만 개의 bead에 붙여진 tag을 분석하기 때문에 매우 빠르다는 장점을 가지고 있다. 최근에는 메머드의 유전체 분석에 사용되기도 하였다 [5].

이외에도 영국의 Solexa사에서는 192Kb의 짧은 인간 유전체 염기서열 분석을 슬라이드위에서 할 수 있는 방법을 개발하였고, Harvard University의 Church교수는 Agencourt Bioscience사를 설립하여 ligation에 의한 sequencing으로 매일 2억 개의 염기서열을 분석할 수 있는 방법을 개발하였는데, 최근에는 매일 30억 개의 염기서열 분석이 가능하다. 또한 Microchip Biotechnology사나 NimbleGen Systems, LI-COR, Network Biosystems, VisiGen Biotechnology 사 등에서 microchip 에 기반한 기술들을 개발하고 있다 [6].

2) SNPlex

대용량 염기서열 분석 플랫폼 중에 전기영동의 원리를 이용한 방법으로 ABI사가 개발한 SNPlex는 48개의 SNP을 하나의 capillary에서 분석할 수 있다. 분석과정을 간단히 설명하면, 이미 알려진 두 가지 genotype에 대해 각각 서로 다른 "ZipCode"가 달린 linker를 primer와 연결한 후에 PCR 증폭을 한다. 이때 각각의 ZipCode는 서로 다른 genotype을 지정하기 때문에 전기영동상 위치를 파악함으로써 genotype calling을 할 수 있게 된다.

이러한 방법으로는 알려진 SNP에 대한 validation 또는 screening에 적합한데, 하나의 capillary당 48개만 분석할 수 있으며, 총 96개의 genotyping이 가능하다. 하지만

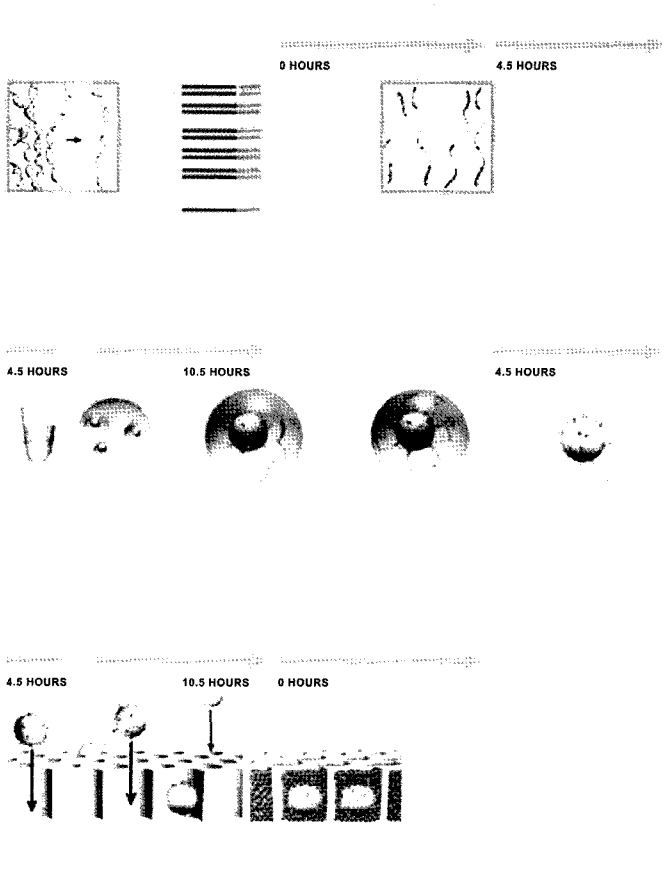

Figure 1. Sequence analysis strategy of 454 life sciences.
(cited from website of 454 Life Sciences, http://www.454.com)

PCR과 전기영동, 그리고 각각의 염기서열에 최적화된 조건을 구하기까지 시간과 노력이 들고 각각 염기서열마다 조건이 다를 수 있어 전체 genome에 대한 분석에는 어려움이 따를 수 있다.

3) GeneChip

앞서 설명한 SNP genotyping 방법에 따른 분류에 의하면 Affymetrix에서 개발한 GeneChip은 염기서열의 상보성에 의해 hybridization이 일어나는가에 따라 genotyping을 할 수 있다 (Figure 2). 최근에 개발된 GeneChip에는 약 50만개의 genotype을 분석할 수 있다. 이들 50만개의 SNP의 선정은 처음 48명의 시료로부터 2.2M SNP을 총 2천5백만 개의 genotype 분석으로 시작하였다. 48명은 각각 코카시아인종, 아프리카인종 및 아시아인종 각 16명으로 구성된 HapMap을 위해 사용한 sample들이다. 첫 번째 데이터로부터 65만개의 후보 SNP을 고른 후에 다시 400명의 시료 (270명의 HapMap sample 포함) 를 이용하여 두 번째 선택을 하였다. 그 방법으로는 Hardy-Weinberg rule과 Mendelian error, 그리고 재현성을 분석하였고, 각 spot별로 call rate를 분석하였다. 최종적으로 50만개의 SNP을 선정하는데 이는 Broad institute에서 분석한 linkage disequilibrium 과 HapMap data를 이용하였다.

실험과정을 간단히 설명하면 Affymetrix 사에서 제공하는 그림 2에 표시된 바와 같이 먼저 genomic DNA를 각각 NspI 또는 StyI 제한효소로 자르고 여기에 linker DNA를 ligation 하여 universal PCR이 가능하도록 한다. 이들 시료를 PCR로 증폭하고 다시 fragmentation 한 후에 end-labeling한다. 이후에 GeneChip에 hybridization 하여 각 spot에서 signal을 분석한다.

이러한 방법의 장점은 50만개의 SNP을 r^2

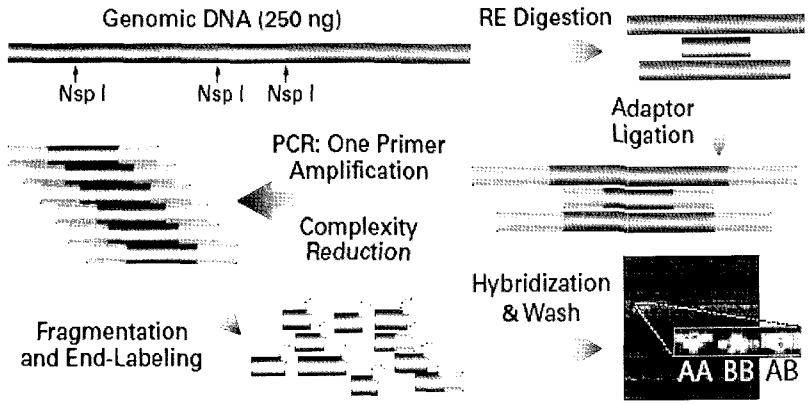

Figure 2. Scheme of SNP genotyping using GeneChip.
(cited from website of Affymetrix, http://www.affymetrix.com)

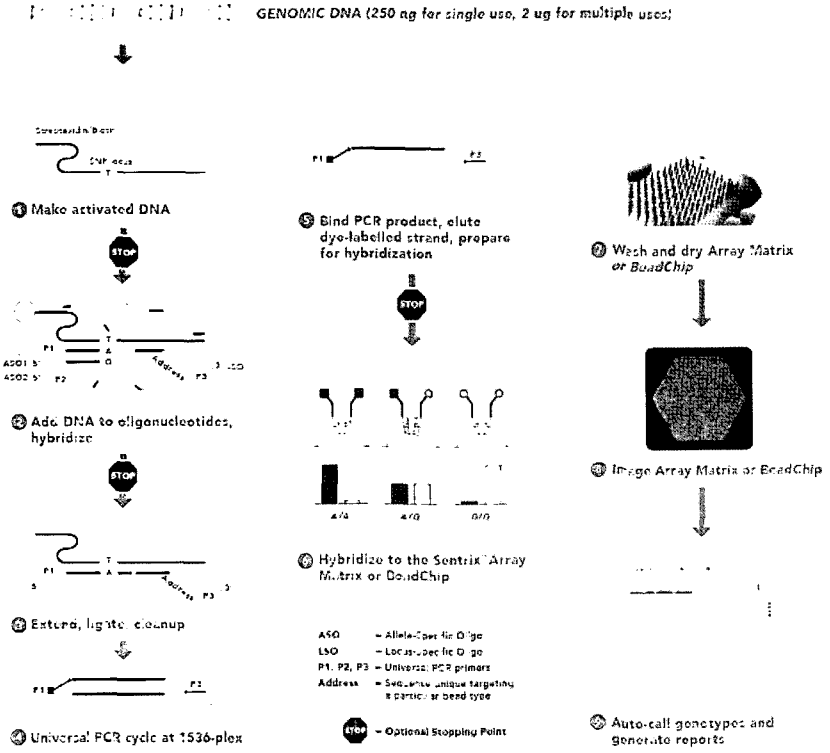

Figure 3. Workflow of SNP genotyping using BeadArray.
(cited from website of Illumina, http://www.illumina.com/)

>0.8일 경우 call rate 81%이상에서 분석이 가능하다. 또한 아시아인종의 call rate 나 coverage도 비교적 코카시안과 유사하다. 이는 HapMap data로부터 Broad institute에서 유전자 및 SNP을 선정한 결과이기 때문이다. 초기에 10K로부터 시작하여 차츰 집적도와 coverage를 높이고 있으며, 최근에는 100K로부터 500K로 향상시키면서 spot의 수도 SNP당 24개로 줄이고 spot의 크기도 5micron으로 줄였으며, pixel size도 0.7micron으로 줄였다.

하지만 다른 유사한 플랫폼에 비하여 call rate가 상대적으로 낮으며, PCR amplification step에서 발생할 수 있는 uneven amplification의 문제를 가지고 있다. 따라서 이러한 점을 고려하여 데이터 분석을 수행하여야 할 것이다. 또한 비교적 정해진 플랫폼을 사용하여야 하기 때문에 custom spotting이나 원하는 subset에 대한 접근이 상대적으로 떨어진다고 할 수 있다. Affymetrix가 microarray 시장에서 차지하는 비중을 생각해 볼 때 미국내 많은 기관과 회사에서 사용 중이며, 기존의 데이터와 호환을 위해 많이 사용 중이다.

4) BeadArray

Illumina사에서 개발한 BeadArray는 SNP genotyping원리중 hybridization과 extension and ligation 두가지 단계에 의해 genotyping이 된다 (Figure 3). 따라서 두 단계의 선택과정을 거치기 때문에 한 번 더 genotype을 체크할 수 있다는 장점이 있다. 또한 bead에서 반응이 이루어지기 때문에 원하는 종류의 SNP genotype을 분석할 수 있어서 유사한 방법에 비해 flexibility가 높다.

비교적 간단하고 대규모 scale로 확장이 가능하기 때문에 자동화가 용이하며, 데이터 분석에서도 LIMS의 적용이 가능하다. GeneChip의 경우와 유사하게 genomic DNA가 250-750 ng 정도로 적게 사용된다. 이는 처음에 whole genome amplification을 통해 대규모로 양을 늘릴 수 있기 때문이다. 하지만 GeneChip과 달리 hybridization 후에 PCR amplification을 하기 때문에 uneven amplification의 문제는 적다고 생각된다.

구체적인 내용을 보면 bead, 즉 spot의 크기는 3 micron정도이며, 각 bead는 5 micron 가량 떨어져 있다. 약 1.5 mm bundle에 대략 50,000 features 정도가 포함되어 있고, 최근에는 하나의 bundle로 1,536개의 SNP을 genotype할 수 있도록 되어 있다. 전체적으로는 317,503개의 SNP loci에 대한 분석이 가능한데 call rate는 pairwise $r^2 > 0.8$에서 99.93% 정도로 보고하고 있다.

현재 전 세계적으로 whole genome 대용량 SNP 분석법으로 GeneChip과 BeadArray이 가장 많이 사용되고 있다. 미국 MIT대학의 Broad institute에서는 Affymetrix의 GeneChip을 주로 사용하고 있으며, 반면에 Harvard대학의 경우 Illumina BeadArray를 사용하고 있다. 미국 국립보건원 산하 NCI에서 2006년 2월에 발표한 Cancer Genetic Markers of Susceptibility (CGEMS) 계획은 전립선암과 유방암에 대한 감수성 유전인자를 찾으려는 것으로 3년간 천4백만불을

투입하여 whole genome에서 검색하려고 한다. 먼저 대조군을 포함하여 전립선암으로 진단된 환자의 2,500명의 시료를 BeadArray 기술을 이용하여 30여만 개의 SNP 분석을 시작하고 있다. 이외에도 African-American Cohort study 및 각종 유전체 코호트 사업에 각종 microarray-based whole genome SNP analysis가 시작되고 있는 실정이다.

결 론

이제까지 살펴본 바와 같이 다양한 genotyping 방법이 개발되어 있으며, 앞으로도 새로운 염기서열 분석 방법이 꾸준히 개발될 전망이다. 이들은 처리용량에 따라 나누어 질 수 있으며, 유사한 플랫폼 중에서도 정해진 SNP만을 분석하기도 하지만 연구자가 원하는 SNP 을 분석할 수 있는 유연성이 높은 기술도 있다. 최대의 분석 효율을 얻기 위해서는 정해진 format을 사용하는 것이 좋으며, 사용자가 원하는 SNP을 모든 플랫폼에서 동일한 효율로 얻을 수는 없다. 즉, 효율 또는 비용과 정보의 유용성을 적당히 거래할 필요가 있다. 따라서 정해진 SNP에 대해 최대의 효율로 분석하고, 이중에서 유용한 SNP에 대해 동일한 플랫폼에서 subset을 가지고 분석하는 것이 효과적인 분석전략이라고 할 수 있다.

대부분의 study design에서 제한요소가 되는 것은 비용이다. 하지만 동일한 자원으로 좋은 결과를 얻기 위해서는 study design을 정교하게 하고 좋은 시료의 질과 임상정보를 확보하는 것이 비용을 줄일 수 있는 방법이기도 하다. 마지막으로 간과하기 쉬운 분야는 데이터의 관리와 데이터베이스이며, 우리가 예상할 수 없을 정도의 분량에 해당하는 유전자 정보와 임상정보 등을 다룰 수 있는 자동화된 체계가 필요하다. 여기에는 시료 및 분석결과, 분석의 정확도 등에 대한 분석을 추적해서 항상 genotyping 중간에도 이를 feedback해 줄 수 있는 시스템이 보완되어야 한다.

미국을 중심으로 시작되는 유전체 코호트 사업들을 보면 전체 3-5년간의 유전체 분석 기간을 설정하고 첫 해에는 각종 유전체 분석 시스템, 특히 chip-based whole genome analysis에 대한 평가를 수행하고 있다. 이는 향후 3년동안에 걸친 유전체 분석 작업의 자동화, 공정화, 데이터분석에 이르는 전 과정을 미리 테스트해 보는 작업들이 포함된다.

앞으로 계획하는 유전체 코호트 사업에서는 2-3년 앞서 나가는 대규모 코호트 사업들을 벤치마킹하는 것이 중요하다. 현재 유전체 분석에서 기술적인 검토를 통해 많은 기관에서 BeadArray를 채택하는 경향을 볼 수 있다. 특히 한국인 유전체에 대한 대용량 분석을 통하여 동일한 시료 및 한국인 pilot 실험을 통하여 적절한 플랫폼을 조기에 결정하고, 이를 기반으로 대용량 분석 플랫폼을 구축하기 위한 투자가 필요하다.

참고문헌

1. Kruglyak L. Power tools for human genetics. *Nat Genetics* 2005; 37(12): 1299-1300
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. *Nature* 2006; 444(7118): 444-454
3. Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, Sinha S, Moore JL, Jagadeeswaran P, Zhao W, Ning G, Makalowska I, McKeigue PM, O'donnell D, Kittles R, Parra EJ, Mangini NJ, Grunwald DJ, Shriver MD, Canfield VA, Cheng KC. SLC24A5, A putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 2005; 310(5755): 1782-1786
4. Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, Van Gestel J, Timmerman P, Zhu M, Lee E, Williams P, de Chaffoy D, Huitric E, Hoffner S, Cambau E, Truffot-Pernot C, Lounis N, Jarlier V. A Diarylquinoline drug active on the ATP synthase of mycobacterium tuberculosis. *Science* 2005; 307(5707): 223-227
5. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 2006; 311(5759): 392-394
6. Service RF. Gene sequencing: The race for the $1000 genome. *Science* 2006; 311(5767): 1544-1546