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 throughputs 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 medicine will focus on the genetic susceptibility of individuals, we need to find genetic variations of each person by efficient genotyping methods.

Key words: Genotype, Sequence analysis, DNA, Microarray analysis, Single nucleotide polymorphism
유전자 코호트 연구를 위한 대용량 염기서열 분석

1. SNP 분석기술 현황

개인간의 유전자 다형성중의 일부는 SNP로 설명할 수 있다. Size variation과 같은 다른 형태의 유전자 다양성에 비해 고르게 분포하며, 길다른 높이 유전자 마커로의 유용성이 높다. 하지만 기존의 전통적인 해석은 당사자에 비해 다양성이 부족하기 때문에 많은 수의 SNP 염기서열 분석이 필요하다. SNP의 장점은 염기서열 분석을 공정하게 대용량 초고속 분석이 가능하다는 것이다. 즉, 특정한 염기서열을 대량으

정하기 위해서는 연구를 개발한 방법으로서 동시에 SNP를 대량으로 분석할 수 있게 한다. 현재 보고된 SNP는 3개만 가능하며, 이들 연구인 중간에서의 효율적인 염기서열 분석법이 필요하다. 본 논문에서는 현재까지 보고된 각각 genotyping methods를 분석하고, 효율적인 대용량 고유분석법에 대하여 설명하고자 한다.

Table 1. Ideal method for high throughputs genotyping

<table>
<thead>
<tr>
<th>genotype method</th>
<th>Affymetrix</th>
<th>Illumina</th>
<th>100K GeneChip</th>
<th>MassARRAY</th>
<th>Applied Biosystems</th>
<th>Sequenom</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP markers</td>
<td>10,000</td>
<td>10,000</td>
<td>10,000</td>
<td>10,000</td>
<td>10,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Selection criteria</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Multiplexing</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Cost</td>
<td>High</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

2. 분석의 범위에 따른 genotyping 방법의 선택

각각의 SNP genotyping 방법에 따라 분석 가능한 SNP의 수와 효율이 다르다. 가장 먼저 본인이 분석하고자 하는 SNP의 개수와 함께, 시료의 숫자를 정하고 이로부터 가장 적합한 분석법을 찾아야 한다. 즉, 10

개 이상의 SNP에 대해 수천 개의 샘플을 분석하기 위해서는 Applied Biosystems의 TaqMan real-time PCR 방법이 제일 유용하다. 이와 Sequenom의 MassARRAY 시스템도 1,000개 이내의 샘플에서 SNP를 분석할 때에는 효율적이다. 또한 기존에 사용되는 각종 genotyping 방법들은, 즉 pyrosequencing이나 sequencing-by-synthesis에 의한 방법들은 100개 이내의 샘플을 위해서도 모두 유용하다고 할 수 있다. MegAllele과 같은 방법은 10,000개의 SNP를 분석할 수 있으며, Affymetrix의 경우 500만 개의 SNP를 동시에 분석할 수 있는 Affymetrix Human Mapping 500K Array Set를 제공하고 있다. 이와 같은 방법들은 전체 유전자 숫자에서 SNP 분석을 가능하게 하는데, 최근에는 Illumina에서 개발한 BeadArray technology를 이용하여 많은 기관에서도 사용하고 있는 실림이다. Table 2에 서 정리한 바와 같이 각각의 genotyping 방법들의 특징과 장단점을 이해하고 시료와 대상 SNP의 숫자를 고려하는 것이 매우 중요하다.

AB1사의 7900 HT는 384개의 TaqMan assays를 30분 만에 처리하거나 48개의 384 well plate, 즉 18,432개 샘플을 블록에 처리할 수 있다. 반면에 Sequenom사의 경우 고해 용량 AutoFux mass spectrometer을 이용 하면, 20개의 384-well chip (7,680 samples)들을 분석할 수 있다. Pyrosequencing을 사용하면 96개의 샘플 을 10분 만에 처리하는데 이를 위해 PCR를 해야 하는 번거로움이 있다.

3. 기존의 Database를 이용한 분석


이들의 전략은 두 가지로 요약할 수 있는데, 첫 번째는 HapMap과 같은 유용한 DB를 이용하여 전단 genotype을 빠르게 찾을 수 있었다는 것이고, 두 번째는 Zebra fish와 같이 간편한 모델동물들을 이용하였다는 것이다. 모든 진화와 생리학적에 적합한 모델동물을 모두 갖는 것이 이상일 것을 이야기하면서, 연구자가 원하는 현상을 적절한 모델동물을 활용하는 것은 수용할 전략중 하나이라고 할 수 있었다.

모델동물과 함께 적절한 인구집단을 사용하는 것도 생각할 수 있으며, 이는 가계 분석이 그로부터 추출한 연구에서 가능 할 것이다. 이는 동물모델이 더 수수한 샘플들이 될 수 있으며, 바로 인간의 질환에 적용할 수 있다. 앞서 예기한 deCode Genetics의 전략으로, 최근 연구집단에 대한 유전자형 분석으로 STR marker 분석으로 질환유전자를 찾고자 할 때 fine mapping의 방법으로, SNP genotyping을 사용하고 있다. 또한 가계분석을 통한 연구성료를 위해서 SNP genotyping과 STR 분석을 병행하여 각 각의 장단점을 적절히 활용하고 있다. 따라서 유전자 목록 분석의 경우에도 SNP genotyping에 다른 marker들의 분석 방법이나 샘플들에 대한 인자를 같이 고려해야 할 것으로 판단된다.

4. 모델동물의 활용

모델동물 자체에서 질환관련 SNP를 분석하는 경우에는 자체의 염기서열에 대한 정보가 필요하다. 최근 염기서열 분석의 결과로 각종 모델동물들의 전체 유전체 염기서열이 밝히지고 있으나, SNP에 대한 자료는 인간에 비해 보잘 것 없다. 즉, 생쥐의 경우 약 6만개의 SNP는 보고되어 있고, 고마우치의 경우 친가, 소의 경우 2만 개가, 그리고 뱀의 경우 백칠십만 개의 SNP는 dbSNP에 보고되어 있다. 최근 염기서열 분석이 증가하고 있는 이들 모델동물이나 작물의 경우 더 많은 SNP가 앞으로 보고될 것이고, 이를 이용하기에 이용하는 것이 필요하다.

5. Genotyping 비용

이론의 genotyping 분석은 대부분의 대형 폐쇄방식으로 고가의 기기와에 대한 투자가 필요하다. 하지만 전반적으로 개별 유전자의 형 분석은 이용되는 비율이 하락한다. 비용의 산정을 위해 고려해야 할 것으로는 각 SNP에 대한 비용이 아니라 sample에 대한 비용이 얼마나 드는가에 대한 것이 다. 즉, 필요에 의해 sample의 수를 늘리아 하는 경우 추가 비용을 계산하여야 하는 경우가 많다. 또한 높은 재현성을 가진 방법은 통계적 처리를 위한 실험 반복의 횟수가 줄어들면서 전체 비용을 줄일 수 있다.

최근 미국 버지니아주의 Bioinformatics학자들에서 분석한 결과에 의하면 보고된 SNP 전체의 40%가뿐만 일부 장관 또는 기관의 중앙지원시설을 통해 분석하였다 한다. 이는 비용을 줄이기 위한 방법이라고 해석할 수 있다. 예를 들어 영국 Oxford에 있는 Wellcome Trust Center에서는 Illumina의 beadArray를 유전체 검색법으로 사용하고, Sequenom사의 MassARRAY를 fine mapping의 방법으로 사용한다. 각각 3-4억 원의 기 초투자가 필요한 장관이다. 미국 보스턴의 MIT대학내 Broad Institute에서는 Sequenom과 Affymetrix platform을 사용한 다. Affymetrix에 비해 Illumina의 array는 자 신이 원하는 SNP를 선택할 수 있기 때문에 장점인 SNP에 점검하기보다 자신의 SNP list를 가진 경우 유리할 수 있다.

<table>
<thead>
<tr>
<th>SNP No.</th>
<th>Cost (Genotype)</th>
<th>Cost per 1,000 sample ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5~10</td>
<td>0.60</td>
<td>6,000</td>
</tr>
<tr>
<td>48~96</td>
<td>0.25~0.30</td>
<td>364~1,556</td>
</tr>
<tr>
<td>30,000~50,000</td>
<td>0.008~0.15</td>
<td>57,600~122,800</td>
</tr>
<tr>
<td>(defined format)</td>
<td></td>
<td>400,000~800,000</td>
</tr>
<tr>
<td>10,000~20,000</td>
<td>0.03</td>
<td>&gt;250,000</td>
</tr>
</tbody>
</table>

6. 각종 대형 genotyping법인 비교

1) 대형량 염기서열 분석

2003년 9월 Venter 박사는 J. Craig Venter Science Foundation을 통해 인간유전자 염기서열 분석을 천천히 가능하게 할 수 있는 기술을 개발하는 사업에 50만불을 수여하였다 한다. 이에 미국 California주 산타로나리에 있는 X Prize Foundation에서는 5백만 불 내지 2천만 본의 상금을 제도하였다. 2004년에는 미국 NIH의 국립유전자연구소에서 Collins박사는 포유동물 크기가의 유전체를 기초에는 10만 본, 국적적으로는 천천히 분석할 수 있는 기술 개발을 위해 7천만 본, 즉 700억 원의 연구비를 투자하였다고 공표하였다. 이러한
한 노력에 힘입어 대용량 초고속 분석법
들이 소개되고 있다.

대표적으로 454 Life Science에서는 bead sequencing법으로 작은 크기의 염기서열을 빠르게 pyrosequencing으로 분석하는 시스템을 구축하였다(Figure 1). 이들은 Mycoplasma genitalium의 2킬로바이트의 염기서열을 99.4%의 정확도를 가지고 분석하여 발표하였으며, 2.1Mb 유전체를 가진 M. tuberculosis를 분석하여 2005년 Science지에 발표하였다 [4]. Pyrosequencing의 문제로 100bp 정도의 짧은 염기서열을 분석하지만 24시간 내에 수십만 개의 bead에 붙여진 tag을 분석하기 때문에 매우 빠르다는 장점이 가지고 있다. 최근에는 메

드의 유전체 분석에 사용되기도 하였다 [5].

이외에도 영국의 Solexa에서는 192Kb
의 짧은 인간 유전체 염기서열 분석을 슬
라이드위에서 할 수 있는 방법을 개발하
고, Harvard University의 Church 교수는
Agencourt Bioscience사를 설립하여 ligation
에 의한 sequencing으로 메일 2억 개의 염
기서열을 분석할 수 있는 방법을 개발하
는데, 최근에는 메일 30억 개의 염기서
열 분석이 가능하다. 또한 Microchip Biote
chnology사나 NimbleGen Systems, LL-COR,
Network Biosystems, VisiGen Biotechnology
사 등에서 microchip에 기반한 기술들을
개발하고 있다 [6].

2) SNPlex

DNA 양가 염기서열 분석 플랫폼 중에 전기
영동의 원리를 이용한 방법으로 ABI사가 개
발한 SNPlex는 48개의 SNP를 하나의
capillary에서 분석할 수 있다. 분석과정은
간단히 설명하면, 이미 알려진 두 가지
genotype을 분석할 수 있는 “ZipCode”
가 담긴 primer를 primer와 연결한 후에 PCR
증폭을 한다. 이후 각각의 ZipCode는 서로
different genotype을 구분하기 때문에 전기영동
동기 분석을 하여서 genotype calling
을 할 수 있게 된다. 이

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

PCR과 전기영동, 그리고 각각의 염기서열
에 최적화된 조건을 구하기까지 시간과
노력이 들고 각각 염기서열마다 조건이
다를 수 있어 전체 genome에 대한 분석에
는 어려움이 따른다. 3) GeneChip

앞서 설명한 SNP genotyping 방법에 따른
분류에 의하면 Affymetrix에서 개발한
GeneChip은 염기서열의 상호성에 의해
hybridization이 일어나는가에 따라 genotyping
을 할 수 있다(Figure 2). 최근에 개발
된 GeneChip에는 약 50만개의 genotype을
분석할 수 있다. 이들 50만개의 SNP의 선
정은 처음 48명의 시료로 2.2M SNP를
총 2천5백만 개의 genotype 분석으로 시작
하였다. 48명은 각각 코카이아인종, 아프라
리안종 및 아시아인종 각 16명으로 구성
된 HapMap을 위해 사용한 sample들이다.

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

이러한 방법의 장점은 50만개의 SNP를
을 고른 후에 다시 400만의 시료 (270만의
HapMap sample 포함) 를 이용하여 두 번째
실험을 하였다. 그 방법으로는 Hardy-Win
eberg rule과 Mendelian error, 그리고 재현
성은 분석하였고, 각 spot별로 call rate를 분
석하였다. 최종적으로 50만개의 SNP를 선
정하는데 이는 Broad institute에서 분석한
linkage disequilibrium과 HapMap data를 이
용하였다.

실험결과를 간단히 설명하면 Affymetrix
사에서 제공하는 그림 2에 표시된 바와 같
이 먼저 genomic DNA를 각각 NspI 또는
StyI 제한효소로 자르고 여기에 linker DNA
를 ligation 하여 universal PCR이 가능하도
록 한다. 이들 시료를 PCR 로 증폭하고 다
시 fragmentation 된 후에 end-labeling 한다.
이후에 GeneChip의 hybridization 하여 각
spot에서 signal을 분석한다.

이러한 방법의 장점은 50만개의 SNP를 r
 callable rate 81%.

HapMap data

Affymetrix and Illumina.

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 고가시안과 유사하다. 이는 HapMap data로부터 Broad institute에서 유전자 및 SNP를 선정한 결과이기 때문이 다. 초기에 10K로부터 시작하여 차츰 점적도와 coverage를 높이고 있으며, 최근에는 100K로부터 500K로 향상시키면서 spot의 수도 SNP당 24개로 줄이고 spot 크기도 5 micron으로 줄었으며, pixel size도 0.7 micron으로 줄었다.

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 저잡된 플랫폼을 사용하여야 하기 때문에 custom spotting이나 원하는 subset에 대한 접근이 상대적으로 벌어진다고 할 수 있다. Affymetrix가 microarray 시장에서 차지하는 비율을 생각해 볼 때 미국내 많은 기관과 회사에서 사용 중이며, 기존의 데이터와 호환을 위해 많이 사용 중이다.

4) BeadArray

Illumina에서 개발한 BeadArray는 SNP genotyping 원리와 hybridization가 extension and ligation 두가지 단계에 의해 genotyping이 된다 (Figure 3). 따라서 두 단계의 선택 과정을 거치기 때문에 한 번 더 genotype을 체크할 수 있다는 장점이 있다. 또한 bead에서 반응이 이루어지기 때문에 동유의 SNP genotype은 분석할 수 있어서 유사한 방법에 비해 flexibility가 높다.

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

구체적인 내용을 보면 bead, 즉 spot의 크기는 3 micron 정도이며, 각 bead는 5 micron 가량 범위이다. 약 1.5 mm bundle에 대해 50,000 features 정도가 포함되어 있고, 최근에는 하나의 bundle로 1,536개의 SNP를 genotype할 수 있도록 되어 있다. 전체적으로는 317,503개의 SNP locus에 대한 분석이 가능하나 call rate는 pairwise r > 0.8에서 99.93% 정도로 보고하고 있다.

현재 전체적으로 whole genome 대용량 SNP 분석법으로 GenChip와 BeadArray가 가장 많이 사용되고 있다. 미국 MIT대학의 Broad institute에서는 Affymetrix의 GenChip을 주로 사용하고 있으며, 반면에 Harvard대학의 경우 Illumina BeadArray를 사용하고 있다. 미국 국립보건소 산하 NCI에서 2006년 2월에 발표한 Cancer Genetic Markers of Susceptibility (CGEMS) 계획은 전산산업과 유전자에 대한 감수성 유전자의 검출을 para 그리고며 3년간 4백만불을
유전체 코호트 연구를 위한 대용량 엑시얼 분석


 nadzieję. 그 후에, 유전체 분석을 위한 시스템이 보완되어야 한다. 미국을 중심으로 시작되는 유전체 코호트 사업들을 보면, 3-5년간의 유전체 분석 기간을 설정하고, 해에는 각 유전체 분석 시스템, 특히 chip-based whole genome analysis에 대한 평가를 수행하고 있다. 이는 후반 3년 동안에 전반 유전체 분석 사업의 자동화, 공정화, 데이터 분석에 이르는 전 과정을 미리 테스트해 보는 작업이 포함된다. 앞으로 제시하는 유전체 코호트 사업에서 서는 3-4년 앞서 나가는 대규모 코호트 사 업들을 배치마킹하는 것이 중요하다. 현재 유전체 분석에서 기술적인 점을 통해 많은 기관에서 BeadArray을 체계화하는 경향을 볼 수 있다. 특히 한국인 유전체에 대한 대용량 분석을 통하여 동일한 시로 및 한국인 pilot 실험을 통하여 적절한 플랫 폰 조기에 결코하고, 이를 기반으로 대 용량 분석 플랫폼을 구축하기 위한 툴가 필요하다.

**참고문헌**


