韓蠶學誌 42(2), 126~141 (2000)
Korean J. Seric. Sci.
< 總 說 >

유전자 및 유전체 연구 기술과 동향

이진성 · 김기환[1] · 서동상[1] · 강석우[2] · 황재삼[2],*
코아바이오 생명과학연구소, [1]성균관대학교 유전공학과, [2]농촌진흥청 농업과학기술원

Trend and Technology of Gene and Genome Research

Jin-Sung Lee, Ki-Hwan Kim[1], Dong-Sang Suh[1], Seok Woo Kang[2] and Jae-Sam Hwang[2],*
CoreBio Institute of Life Science & Biotechnology, Corebio System Co. LtD, Seoul, Korea
[1]Department of Genetic Engineering, Sungkyunkwan University, Suwon, Korea
[2]Depatment of sericulture & Entomology, National Institute of Agricultural & Technology, RDA, Suwon, Korea

ABSTRACT

A major step towards understanding of the genetic basis of an organism is the complete sequence determination of all genes in target genome. The nucleotide sequence encoded in the genome contains the information that specifies the amino acid sequence of every protein and functional RNA molecule. In principle, it will be possible to identify every protein resposible for the structure and function of the body of the target organism. The pattern of expression in different cell types will specify where and when each protein is used. The amino acid sequence of the proteins encoded by each gene will be derived from the conceptional translation of the nucleotide sequence. Comparison of these sequences with those of known proteins, whose sequences are sorted in database, will suggest an approximate function for many proteins. This mini review describes the development of new sequencing methods and the optimization of sequencing strategies for whole genome, various cDNA and genomic analysis.

서 론

분자생물학 용어인 central dogma란 어떻게 유전자가 암호화된 DNA 염기서열을 mRNA로 전사하며 그리고 그것이 기능을 갖는 단백질로 어떻게 번역되는 가를 정의하는 것이다. 이것은 처음에 프랜시스 클릭이 1957년에 최초로 제안하여 현재 중요한 역사적 발견으로 인정되고 있다. 1966년 전체 유전자의 암호(genetic code)가 Korana등에 의해서 발견되어 DNA 서열로부터 단백질 서열을 추론할 수 있게 되었다. 10년 후에 태동된 DNA 염기서열 분석법은 Maxam-Gilbert 그룹과 Sanger(Maxam & Gilbert, 1977; Sanger *et al.*, 1977)에 의해서 소개되어 16.5 kb의 인간 미토콘드리아 유전체(genome)와 40 kb로 구성된 람다 박테리오 파지의 전체 염기서열이 결정되는 데 기여하였다(Anderson *et al.*, 1981; Sanger *et al.*, 1982). 이후로 염기서열 분석 기술, 특히 Sanger 등이 고안한 효소적 체인 종료 방법(enzymatic chain termination)이 더욱 더 개량되어 염기서열 분석 자동화의 중요한 기술로 채택되었다. 대규모 염기서열의 극적인 증가는 각기 다른 고유한 특징을 지닌 생물의 유전체를 해석할 수 있게끔 하는 중요한 원동력이 되었다.

본 논의는 20세기의 총아라 불리는 전체 유전체 해석에 집중된 다양한 DNA 해석기술의 세계적 흐름과 누에의 유전자 연구 및 유전체 연구의 기초 자료로 제시하고자 한다.

1. 유전체 계획

인간 유전체 계획(human genome project, HGP)은 15년 동안에 걸쳐 전체 인간 유전체의 물리지도 작성과 전체 염기서열 해독을 완료한다는 취지로 1990년에 시작되었다. 또한 이 야심찬 계획은 기능 유전체학이라는 새로운 학문을 태동시켰으며 궁극적으로 수많은 질병에 고통 받고 있는 인류에게 희망찬 비전을 제시하였다. 현재 다양한 생명활동을 영위하는 생명체들이 완전 염기서열 결정을 위한 목적 생물로 정해졌으며 따라서 앞으로 이들 다양한 생명체의 유전체 해독 결과물은 인간 유전자의 기능을 이해하는데 많은 정보를 제공할 것이다.

출아 효모(budding yeast)는 1996년에 국가간 협력체

(International consortium)에 의해서 완전한 유전체 염기서열이 해독되었다(Goffeau *et al.*, 1997). 출아 효모는 작은 유전체를 가지고 있으며, 염색체와 반복 DNA 서열이 적고 intron이 거의 없다는 특징을 갖는다(Oliver *et al.*, 1996). 또한 단세포계이며 정의된 배지(defined medium)에서 생육하는 특징을 가지며 반수체 및 배수체 상태로 생육이 가능하며 유전자의 결실을 연구할 수 있는 방법이 개발되어 있어 기능 유전체 연구에 유용한 모델 생물이 되었다. 아직까지 인간의 모든 유전자가 결정되어 있지 않았지만 출아 효모의 유전체를 분석한 결과 인간과 31%라는 놀라운 유전적 유사성이 있음이 발견되었다(Bostein *et al.*, 1997). 출아 효모 유전체의 가장 놀라운 특징은 이들의 유전체 중에서 70% 이상이 단백질을 암호화는 유전자로 구성되어 있다는 점과 이들 유전자의 4.5%만이 intron을 가지고 있는 것이었다. 최근에 선충의 일종인 *C. elegans*의 유전체 전체의 염기서열이 해독되었으며 인간과는 36%의 유전적 유사성이 있음이 밝혀졌다. 선충이 유전체 계획의 모델 생물로 선택된 이유는 단순 다세포계 생물이라는 것 때문이며 특히, 다세포계의 분화 및 신경계의 기능 연구에 중요한 연구 재료였기 때문이다(Sulston *et al.*, 1992). 또한, 식물 연구의 모델 생물인 *A. thaliana*는 작은 크기의 유전체, 빠른 생육, 자가 수분, 적은 염색체 수 등의 장점으로 전체 염기서열 분석이 완료되었다. 또한, *A. thaliana*는 반복 DNA 서열이 적으며 효모의 2 kb 당 한 개에 비교해서 4.8 kb 당 한 개의 비교적 높은 유전자 밀도를 가지는 특징이 있다. 전체 염기서열이 분석된 최초의 원핵생물은 1.83 Mb의 유전체를 갖는 감기의 원인균인 *H. influenzae*이다(Fleischmann *et al.*, 1995). 이후 지금까지 분석된 가장 작은 유전체(0.58 Mb)를 갖는 *M. genitalium*의 전체 염기서열이 결정되었다(Fraser *et al.*, 1995). 현재까지 18종류의 미생물 유전체가 완전하게 해독되어 발표되었으며(Table 1) 몇몇 종에 대해서는 현재 유전체 분석이 진행 중에 있다.

최근 방대한 유전체 연구에 필요한 기술적 진보를 맞이하고 있다. 예를 들면 물리지도 작성을 위한 최적화된 플라스미드, 개량된 효소 및 DNA 염기서열 분석을 위한 형광염료의 부착 방법, 자동화된 시료의 준비 시스템, 염기서열 결과 분석을 위한 효율적인 컴퓨터 운영체계 등이 개발되었다. 인간 유전체 전체의 염기서열 해독은 이와 같은 기본계획(Rowen *et al.*, 1997)에 따라 시작되었으며 아마도 2003년에 종료될 것이다. 따라서 앞으로 3년 후에는 인간 역사상 가장 위대한 결과물을 맞이 할 것이다. 그러나 이러한 계획은 PE Biosystems에 의해서 전폭적인 지원을 받고 설립한 유전체 벤처기업인 셀레라 지노믹스(Cerera Genomics, USA)의 Venter 등에 의해 추진되는 산탄 유전자 분석법(Shotgun sequencing)에 의해 강력한 도전을 받고 있으며 이들은 2001년에 완전한 인간 유전체를 해독하겠다고 공언하고 있다(Venter *et al.*, 1998).

1) 유전자 발현 꼬리표 (ESTs)

고등생물 유전체의 염기서열은 단백질을 암호화하는 유전자의 일부분만을 포함하고 있다. 인간의 경우 30억 bp의 유전체 중에서 단지 3%만이 단백질을 암호화하는 것으로 알려지고 있다. 그러나 mRNA로부터 역전사하여 얻은 상보성 DNA(Complementary DNA; cDNA)는 단백질을 암호화하는 유전자의 부분 염기서열로써 직접적인 유전자 공학의 핵심 재료가 되고 있다. 따라서 cDNA의 염기서열 해독은 유전체 계획의 중요한 실마리를 제공하였으며 현재 많은 기법들이 소개되고 있다.

1991년 세계 최초로 대규모의 cDNA 염기서열을 응용한 기술이 인간의 뇌로부터 수행되었다(Adams *et al.*, 1991). 이들 부분 염기서열은 특정시기에 특정조직에서 발현되는 유전자들로 정의되기 때문에 ESTs(Expressed Sequence Tags)라는 용어로 정의된다. 이후로 유사한 연구들이 다른 생물의 여러 조직에서 수행되었다. ESTs는 염기서열을 근간으로 하는 다양한 연구 계획에 상당한 양의 정보를 이용 가능하게 해 주고 있으며 이는 GenBank의 dbEST와 같이 재정지원을 받으면서 상업적으로 성공하는 일례를 보여주었다. 특히 다국적회사인 Merck는 공공의 관심과 경쟁의 확보 측면에서 생명공학 I.M.A.G.E 국제 협력체(Lennon *et al.*, 1996)로부터 대량의 cDNA 클론을 분양

Table 1. Examples of eukaryotic organisms that are currently subjected to whole genome sequencing

Organism	Genome size (Mbp)	No. genes	Sequencing completed	dbEST entries (May 1999)
Homo sapiens	3,000	80,000	2001*	1,380,000
Mus musculus	3,000	80,000	2005*	520,000
Drosophila melanogaster	165	12,000	1999*	83,000
Caenorhabditis elegans	100	19,000	1998	73,000
Arabidopsis thaliana	120	21,000	2000*	38,000
Saccharomyces cerevisiae	12	6,000	1996	3,000

*Estimated time for complete sequence.

받아 cDNA 염기서열 분석을 수행해온 워싱턴 대학에 1994년부터 스폰서를 제공하고 있다.

ESTs 염기서열은 새로운 유전자를 동정하는 데 매우 유용한 기법이며 동시에 몇 가지 다른 측면에서도 중요한 기술로 평가받고 있다. cDNA 클론의 임의적인 선별에 의해서 얻어진 염기서열은 목표 조직에 관한 유전자 발현의 수준과 복잡성에 대한 통계학적 통찰력을 제시하여 환경 요인과 조직 특이적인 유전자 발현의 생체 내 영향을 다각적으로 연구 할 수 있게 해 주었다(Okubo *et al.*, 1992). 또한, 상응하는 cDNA 부분의 염색체 위치를 결정하는 물리지도 작성에도 중요한 실마리를 제공하였으며 intron 및 exon(실제로 단백질로 번역되는 유전자 부위)의 경계를 이해하는 데에도 중요한 단서를 제공하였다. 궁극적으로 유전체 계획의 중요한 목적중 하나인 유전체의 전사부위 결정과 유전자 발현 부위의 예측을 가능하게 해 주었으며 인간 질병 유전자의 연구에도 중요한 단서를 제공하여 단일 질병 유전자 다형현상 연구(Single Nucleotide Polymorphisms, SNP; Landegren *et al.*, 1998) 및 질병 유전자 특성 규명에 중요한 재료가 되었다. 따라서 ESTs는 질병 유전자의 위치 클로닝(Positional cloning; Collins, 1995)을 위한 고밀도 물리지도 작성에 촉진제 역할을 할 것이며 궁극적으로 생물학적 분류계를 뛰어 넘어 관심이 있는 유전자의 클로닝을 통한 진화 유전체학(evolutionary genomics) 분야에도 많은 정보를 제공할 것이다(Marra *et al.*, 1998).

2. DNA 염기서열 분석 기술

염기서열 분석의 대규모 계획에 대한 요구의 증가에도 불구하고 20년 전에 개발된 DNA 분석의 기본 기술이 아직까지 사용되고 있을 정도로 이에 관한 기술의 진보는 답보 상태이다. 오히려 많은 시료의 준비를 가능하게 하는 기술(자동화 염기서열 분석기, 로보트 시료 준비기, 최적화된 시약, 생명공학적으로 개량된 효소 및 염료)의 진보가 더 빨리 진행 중에 있다. 특히, 이와 같은 기본적인 염기서열 분석 방법 및 원리에 부가적으로 대규모 염기서열 분석이 가능케 된 것은 소위 효소 중합 연쇄 반응(Polymerase Chain Reation, PCR ; Saiki *et al.*, 1985)이라는 새로운 기법이 도입되면서 부터이다. PCR 기법은 유전체 연구에 관련된 분석, 탐색 등의 DNA 조작기법에 대한 혁명적 방법이 되었다.

1) 표지법과 검출법의 원리

DNA 염기서열 분석을 위한 가장 중요한 기술은 효소적 체인 종료법(Sanger *et al.*, 1977)과 화학적 분해법(Chemical degradation ; Maxam & Gilbert, 1977)이다. 이 두 가지 방법은 아크릴아마이드 겔 상에서 단일 가닥 나선을 연속적으로 분리하는 원리를 응용한 것이다. Maxam-Gilbert의 염기서열 분석법과 비교할 때 Sanger의 체인 종료 기법은 결과의 분석 방법과 해석이 용이한 장점으로 현재 전 세계적으로 사용되고 있다. 현재는 체인 종료법의 기술을 변형하여 (DNA 단편의 표지, 뉴클레오타이드의 개발, 자동화 및 새로운 시약 및 검출법의 개발) 보다 쉽게 응용 가능하게 되었다.

대규모 염기서열을 위한 하나의 중요한 단계는 자동화된 장비의 개발이다. 이 장비는 레이저 빔에 의해서 방출된 형광 염료로 표지된 DNA 단편을 연속적으로 분석하는 전기영동 단계가 결합된 장치이다. 따라서 이와 같은 장비는 분석 시간을 상당히 줄일 뿐만 아니라 동위원소 사용에 따른 위험성을 배제시킬 수 있었다. 현재 두 가지의 원리를 이용한 DNA 자동 분석기가 사용되고 있다. 하나는 DNA를 구성하는 A, T, C, G 4종류의 염기를 한가지의 형광 염료를 사용하여 표지 반응시키는 방법으로 이 방법은 아크릴아마이드 젤 상에서 4개의 lane을 사용하게 된다(One dye-four lane; Ansorge *et al.*, 1986; Prober *et al.*, 1987; Brumbaugh *et al.*, 1988). 둘째는 한번의 반응에 서로 다른 4가지의 형광 염료를 사용하는 방법으로 이것은 아크릴아마이드 젤 상에서 한 개의 lane을 사용하게 된다(Four dye-one lane; Smith *et al.*, 1986). 현재 전자의 방법을 응용한 장비는 Amersham Phamacia Biotech (USA)과 LI-COR (USA)에서 상업적으로 판매되고 있다. 이들 장비에 의해서 나온 결과는 DNA 단편의 일정한 이동도를 주기 때문에 쉽고 정확하게 DNA를 분석할 수 있다. 따라서 다형분석(Polymorphism)과 같은 진단 및 법의학 분야에 주로 사용되고 있다. 그러나 대량 분석 능력은 four dye-one lane 방법에 비해 현저히 떨어지는 것으로 평가되고 있다. 따라서 four dye-one lane 방식이 유전체의 대량 분석을 위해서는 바람직한 방법이다. 이에 기초를 둔 응용 가능한 장비가 Applied Biosystems (USA) 회사에 의해서 ABI373, 310, 377이라는 이름으로 개발되어 현재 한 번의 가동으로 96개의 시료를 분석할 수 있는 최신의 ABI3700까지 생산되어 관련 분야, 특히 유전체 연구에 지대한 공헌을 끼쳤다. ABI 자동 염기서열 분석 장비의 대규모 처리 능력은 다음과 같은 2가지의 다른 원리를 응용 가능하게 해 주었다. 즉 dye primer 방식과 dye terminator 방식이다. 특히, dye terminator 방식은 지금까지 합성 oligo에 의한 연속적 분석법(Primer walking) 과 같이 비용 및 분석 시간의 단점을 해결한 기술로 평가된다. 또한, 이와 같은 기술의 개발은 미세관을 이용한 분석기의 개발, hybridsaction에 의한 새로운 염기서열 분석 기술, mass spectroscopy에 의한 염기서열 분석법 및 pyrosequencing이라는 새로운 기술을 태동시키는 계기를 마련해 주었다.

2) DNA 염기서열 해독에 필요한 효소

DNA 염기서열 분석을 위한 다양한 전략은 다양한 관련 효소, 형광염료 표지법과 같은 신기술의 개발이 토대가 되었다. 현재 다양한 특징을 갖는 DNA 중합효소가 DNA 염기서열 분석에 이용된다. 대장균 DNA 중합효소의 일부분인 klenow 단편(Klenow & Henningsen, 1970)은 DNA 염기서열 분석에 이용된 최초의 효소이다. 이 효소의 단점은 dNTP에 대한 ddNTP의 주형 의존형 discrimination으로 염기서열 반응 시에 각 염기에 대한 단편의 양적 변이가 심하다는 것이었다. 따라서 실제 결과에서 균일한 peak를 보여주지 못하였다. 결국 이것은 DNA 해석에 장애가 되었으며 이러한 단점은 변형된 T7 DNA 중합효소 개발(Tabor & Richardson, 1987)의 촉진제가 되었다. 이 효소는 Mg^{2+}보다 Mn^{2+} 존재 하에서 최적의 활성을 보이며 기존의 klenow보다 우수한 결과를 보여 주어 결국 염기서열 결과의 정확성 및 신빙성을 높인 계기가 되었다.

PCR 기법이 도입되자 *Thermus aquaticus(Taq)*에서 분리된 열에 안정한 DNA 중합효소가 DNA 분석법에 이용되었다(Chien *et al.*, 1976). Sanger법의 원리에 의해서 나타나는 단일가닥 DNA의 합성 원리가 이때부터 PCR 이라는 기법으로 응용되어 'cycle sequencing'이라는 새로운 용어가 등장하게 된 것이다(Innis *et al.*, 1988; Carothers *et al.*, 1989; Murray, 1989). Cycle sequencing 기법은 단순성 및 용이성 때문에 상당히 대중적인 방법이 되었으나 Klenow 효소와 같이 주형으로 ddNTP가 incorporation 되는 효율이 낮아 많은 양의 ddNTP가 요구된다는 단점은 극복하기 어려웠다(Lee *et al.*, 1992; Khurshid & Bexk, 1993).

DNA 서열 분석법의 대량화에 직면하면서 나타난 일부 DNA 중합효소의 ddNTP 낮은 인식능력은 활성 부위내에 존재하는 페닐알라닌의 hydroxyl group 때문인 것으로 밝혀지게 되었다. 원래의 대장균 DNA 중합효소 I과 *Taq* DNA 중합효소는 그들의 활성 부위에 T7 중합효소가 타이로신을 갖는데 반해 페닐알라닌을 갖는다. 따라서 연구자들은 유전공학적으로 대장균 DNA 중합효소와 *Taq* DNA 중합효소의 페닐알라닌을 타이로신으로 치환함으로써 이를 극복하였다. 결국 ddNTP의 효율을 250-8000배 증가시키는 결과를 이끌었다(Tabor & Richardson, 1995). 결국 돌연변이화된 *Taq* DNA 중합효소 F667Y를 이용한 cycle sequencing 결과는 균일한 peak를 보여 주었으며 T7 DNA polymerase와 비교할 때 어떠한 차이도 보여주지 않았다. 또한, 같은 양의 ddNTP를 사용할 수 있게 되면서 자동 염기서열 장비 사용 시에 나타나는 형광 염료의 background 도 줄어드는 효과도 얻을 수 있었다. 따라서 돌연변이로 생산된 *Taq* DNA 중합효소는 DNA 염기서열 분석 기술에 상당한 영향을 주었으며 새로운 염기서열 분석법 및 그 응용에 이용되는 계기가 되었다.

3) DNA 주형의 특징

DNA 염기서열 반응의 성공 여부는 주형 DNA의 형태 및 순도에 크게 의존한다. Sanger의 원리를 토대로 하는 염기서열은 단일나선 DNA를 주형으로 하나의 primer가 신장하여 나오는 DNA 단편을 분석하는 것이다. 여기에는 ssDNA, dsDNA와 같은 주형의 종류에 따라 다양한 방법이 현재 응용되고 있다. T7 DNA 중합효소 또는 klenow를 이용한 주형 ssDNA의 준비는 M13 파지의 감염을 통한 목적 부위의 클로닝에 의해서 얻을 수 있다(Messing *et al.*, 1978). 대장균에 감염시킨 후, M13 파지에 의해서 분비된 단일 가닥 DNA는 배지로부터 수집된다. 그러나 이와 같은 과정은 노동력이 많이 필요하게 되어 유전체 계획에 이용하기에는 많은 장애가 된다(Smith *et al.*, 1990; Zimmermann *et al.*, 1990). 따라서 자동화된 장비가 대규모 염기서열 분석을 위해서 필요하게 되었다. 플라스미드로부터 단일가닥 주형 DNA 생산을 위한 한가지의 대안은 고체상 염기서열 분석 기술이다(Stahl *et al.*, 1988). 이 기법은 염기서열 분석시 주형 DNA로서 PCR 산물을 사용하면서 부터 더욱 개량되었다(Hultman *et al.*, 1989). DNA 염기서열 결정을 위한 주형 DNA는 biotin으로 표지된 프라이머에 의해서 합성되며 이것은 streptoavidin으로 싸여진 magnetic bead상에서 PCR 산물의 고정화가 가능하게 된다. 그리고 나서 주형 DNA의 분리는 알카리에 의해서 얻어질 수 있으며 bead는 자석에 의해서 상등액으로 부터 분리할 수 있게 되었다. 따라서 염기서열을 위한 ssDNA는 상등액에 존재하는 bead로써 분리할 수 있게 된 것이다. 따라서 고체상 염기서열 분석기술은 자동화에 적합한 것이 되었으며, 특히 대규모 염기서열 계획에 유용한 기술이 되었다. 또한, PCR 산물을 ssDNA로 전환하기 위해서 비대칭(Asymmetric) PCR 기법(Gyllensten,1989), exonuclease generated PCR 기법(Higuchi & Ochman, 1989), 및 transcript sequencing(Sarkar & Sommer 1988; Stoflet *et al.*, 1988)이라는 새로운 기술의 개발을 촉진시키는 계기가 되었다.

단일가닥 주형과 달리 이중가닥 주형(플라스미드 혹은 PCR 산물)은 직접적으로 염기서열 분석에 사용할 수 있다. 그러나 낮은 온도의 염기서열 반응에 주형으로서 플라스미드를 사용하기 위해서는 플라스미드를 열 또는 알카리에 의해서 변성을 시켜야 한다(Vieira & Messing, 1982; Chen & Seeburg, 1985). M13 파지와 비교할 때, 플라스미드는 준비과정이 쉬우며 빠르다는 장점이 있으나 ssDNA처럼 정확하고 긴 염기서열 결과를 보여주지 못한다. 주형을 준비하는 시간적 측면에서 PCR 산물은 최선의 대

안이다. 왜냐하면 주형 DNA는 박테리아 혹은 플라크로부터 직접적으로 PCR에 의해서 얻을 수 있기 때문이다. 그러나 PCR 산물도 낮은 온도에서 반응하는 염기서열 분석엔 플라스미드와 같은 전처리가 요구된다(Kusukawa *et al.*, 1990; Gyllensten *et al.*, 1992).

M13 파지, 플라스미드 및 PCR 산물간에는 일반적으로 몇 가지의 차이점이 존재한다. 단일가닥 M13 파지는 오직 한 방향으로만 염기서열을 결정할 수 있으나 일반적으로 2-3 kb의 분석 능력을 갖는 PCR 산물과 비교하면 크기의 제한을 덜 받는다. 긴 PCR 산물은 특정 조건 하에서만 얻을 수 있으며 염기서열 분석에 자주 이용되지 않는다. 플라스미드가 low copy 복제 원점을 갖는 벡터로 사용된다면 약 10 kb 정도의 외래 DNA를 가질 수 있다. 이것은 primer walking 기법 및 short-gun 기법에 의한 틈새 메우기 (gap-filling) 과정에 상당한 잇점을 제공한다. 염기서열 결정 반응 시 PCR 산물의 독특한 특징 중 하나는 polymerase slippage 현상이다(Hoog, 1991). 이와 같은 용어는 *Taq* DNA 중합효소가 12개 이상으로 구성된 특정 염기 stretch(homopolymer)에 정확한 수의 상보적인 염기를 incoporation하는 능력이 떨어진다는 의미에서 정의된 것이다. 각 PCR 회전에서 *Taq* 중합효소가 다른 수의 상보적인 염기를 incoporation 하기 때문에 염기서열 결과의 정확성이 떨어지게 되는 것이다. 이것은 anchor primer (5-oligodT-A/C/G-3)에 의해서 극복할 수는 있지만 3' 끝에 poly(A)-tail을 갖는 PCR-amplified cDNA 클론이나 ESTs의 염기서열을 결정하는데 심각한 장애가 된다(Buess, *et al.*, 1997).

4) 대규모 염기서열 분석

대규모 염기서열 분석을 위해서 어떠한 분석 기술을 선택하는 냐가 효율성에 많은 차이를 낸다. 자동화 시스템에 의한 시료의 준비는 필수적이며 또한, 주형의 준비뿐만 아니라 염기서열 반응 또한 자동화된 장비에 의해서 최적화되어야 한다. 그러나 대규모 염기서열 결정을 위한 자동화 시스템의 체계적, 기술적 개념의 정리가 아직까지는 명료하지 않다. 따라서 대규모 염기서열 계획은 반자동화된 장비의 조직적 통합을 근간으로 해야 한다. 처음에 시약의 pipetting 부터 자동화되어야 하며 이것은 연속적으로 PCR 반응과 염기서열 반응을 위한 자동화로 이어져야 한다. 이 단계에서 상업적으로 이용 가능한 장비는 Biomek 1000과 2000(beckman Instruments)과 Catalyst Labstation(Applied biosystems)이다. 에탄올 침전을 포함하는 다음 단계의 작업은 아직 자동화되어 있지 않다. 또한, 주형의 준비(플라스미드 및 파지)를 위한 자동화 장비가 사용자들에게 인정은 받고 있지만 아직까지 상업적으로 이용 가능하지 못하다. 염기서열 분석을 위한 PCR 산물의 사용은 위와 같은 과정을 일부 보상할 수 있으며 인간 유전체 계획에 산탄 방법이 도입되면서 상당한 자동화가 완비되었다(Venter *et al.*, 1998). 대규모 염기서열 분석에서 클론/플라크-picking을 위한 자동화 장비가 이용되고 있으나 염기서열 분석 단계, 즉 반응된 시료의 loading은 아직도 인간의 손에 의해서 수행되고 있다. 이러한 장애를 극복한 것이 바로 미세관 전기영동 방식을 토대로 한 자동화 장비의 탄생이다. 또한, 자동화 시스템은 이를 운용하기 위해 숙련된 연구자들을 필요로 한다.

화학적 배경을 기반으로 하는 기술 또한 대규모 염기서열 분석을 위한 자동화 장비의 출현에 지대한 공헌을 하였다. Cycle sequencing은 자동화 시료 장비에 의해서 생산된 적은 농도의 주형을 이용 가능하게 해 주었다. 또한, 새로운 효소의 개발은 상당한 기술적 진보를 가능하게 하였다. 고체상 염기서열 분석법은 플라스미드 DNA의 준비 및 염기서열 결정 반응시 ethanol 침전 단계를 생략해 주어 깨끗한 결과를 보여주게 되었다(Hultman *et al.*, 1991). 네 가지의 형광 염료의 사용은 기존의 4개의 tube에서 수행되어온 PCR 반응을 한 개의 tube에서 가능하게 해 주었으며 primer의 표지 반응이 필요 없어 실제로 primer walking 전략에 응용이 가능하게 되었다(Tong & Smith, 1992).

Table 2. A total of 18 different bacteria and archaea (a) have been completely sequenced and published. Among the sequencing strategies applied are shotgun (S), directed sequencing (D) and clone by clone approaches (C)

Organism	Genome size (Mbp)	Method
Haemophilus influenzae	1.83	S
Mycoplasma genitalium	0.58	S
Methanococcus jannaschii (a)	1.66	S
Mycoplsma pneumoniae	0.81	C, D
Synechocystis sp.	3.57	C, S
Methanobacterium thermoautotrophicum (a)	1.75	S
Escherichia coli	4.60	C, S
Helicobactor pylori	1.66	S
Archaeoglobus fulgidus (a)	2.18	S
Borrrelia burgdorferi	1.44	S
Bacillus subtilis	4.20	C, S, D
Mycobacterium tuberculosis	4.40	C, S
Treponema pullidum	1.14	S
Pyrococcus aeolicus (a)	1.80	C, S
Aquifex aeolicus	1.50	S
Chlamydia trachomatis	1.04	S
Rikettsia prowazekii	1.11	S
Helicobactor pylori (J99)	1.64	S

3. 유전체 염기서열 결정

생물의 유전체들은 크기와 복잡성에 상당한 차이가 존재한다. 박테리아는 580 kb도 않 되는 유전체와 약 500개의 유전자를 가지고 생명활동을 영위한다(Table 2). 반면에 인간의 유전체는 세균에 비해서 약 5000배의 크기를 갖으며 대략적으로 5000에서 10만개의 유전자를 가지고 있다(Table1). 더군다나 진핵생물과 원핵생물의 유전체 구성은 상당한 차이를 보인다. 박테리아 유전체는 유전자의 밀도가 상당히 높으며 intron이 거의 존재하지 않는데 비해서 진핵생물의 유전체는 상당한 부위가 비 암호 부위로 구성되어 있으며 대부분의 유전자는 intron을 포함하고 있다.

따라서 유전체 염기서열을 분석하기 위해서는 다각적으로 목적 유전체의 특성을 파악하는 것이 일차적으로 중요하며 다양한 접근 방법을 강구해야 한다. 그러나 일반적으로 모든 유전체를 염기서열 분석에 적합한 단편으로 서브 클로닝 할 수만 있다면 충분한 정확성으로 전체 유전체의 해독이 가능해 진다.

1) 물리지도의 작성

물리지도 작성은 염색체 상에서 큰 단편들의 순서를 결정하는 작업으로 정의할 수 있다. 순서가 결정된 클론은 염기서열 분석에 적당한 크기를 갖는 플라스미드나 파지로 더 작게 만드는 작업의 기초가 된다. 따라서 목적 유전체에 대한 물리지도의 작성은 전체 유전체 해석의 가장 중요한 단계인 것이다.

다양한 종류의 벡터가 물리지도 작성을 위해 사용되고 있다(Fig. 1). 이들 벡터는 다양한 크기의 insert를 가질 수 있게 설계되어 있다. 효모 인공 염색체(YAC)는 대략 250 kb에서 1 Mb의 insert를 가질 수 있으며(Burke *et al.*, 1987),

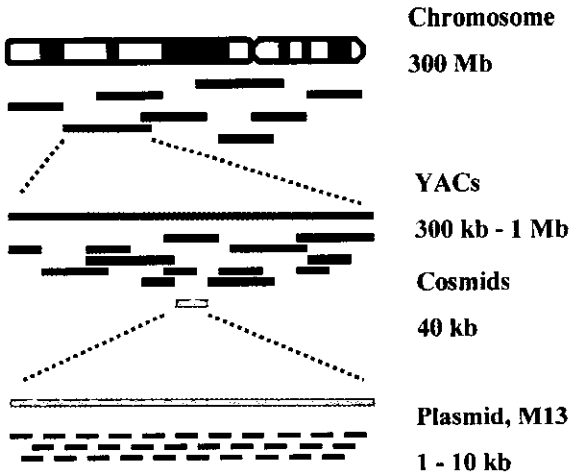

Fig. 1. Examples of vectors used for physical mapping and sequencing and the range of insert sizes that can be harbored by the respective vector.

인간 유전체 대부분을 포함하는 contig(중첩 클론들의 연속적인 set) 작성에 유용하게 사용된다(Cohen *et al.*, 1993). 그러나 chimerism, rearrangment라는 단점을 함께 가지고 있다(Selleri *et al.*, 1992; Chumakov *et al.*, 1995).

코스미드는 람다파지의 cos 부위를 갖는 플라스미드 벡터의 일종이다(Collins & Hohn, 1978). 코스미드로 클로닝된 DNA는 감염을 통해서 대장균으로 효과적으로 도입될 수 있으며 결과적으로 코스미드는 대장균 안에서 플라스미드처럼 유지된다. 코스미드는 40 kb의 평균 insert를 갖는다. 따라서 YAC 클론 사이의 연결부위 클로닝에 주로 이용되며 염기서열 분석용 벡터로도 이용 가능하다. 물론 유전체 일부 부위를 코스미드로 클로닝한 것이 어렵다는 보고도 있고 high copy number라는 것 때문에 다소 불안정하다는 단점도 있으나 플라스미드와 달리 코스미드가 비교적 큰 부위의 유전체 부위를 클로닝 할 수 있다는 최대의 장점으로 유전체를 분석하는 연구자에게는 상당한 잇점을 제공한다(Thierry *et al.*, 1995). 또 다른 벡터가 박테리오파지 P1을 기초로 하여 개발되어 이용되고 있다. 이 P1 벡터는 약 100 kb의 insert를 클로닝 할 수 있는 장점이 있으나 사용하기가 좀 까다롭다는 단점이 있다(Sternberg, 1990).

유전체 연구에 집중적인 연구가 진행되면서 코스미드와 YAC 사이의 insert 크기의 한계를 대체할 수 있고 안정하게 insert를 유지 할 수 있는 벡터의 필요성이 강하게 대두되었다. 이 시점에서 대장균 F factor를 이용하여 15-300 kb의 insert를 클로닝 할 수 있는 박테리아 인공 염색체(BAC)가 개발되었다(Shizuya *et al.*, 1992). F 벡터의 복제는 대장균 내에서 엄격히 조절되며 이를 갖는 BAC은 low copy number로 대장균 내에서 쉽게 유지되기 때문에 잠재적인 재조합 위험성을 배제할 수 있었다. 전기충격법에 의한 고효율 클로닝, 손쉬운 조작, 그리고 안정한 유지성 등으로 BAC은 코스미드 보다 물리지도 작성에 더 적합한 것으로 평가되고 있다. 또한, P1 벡터의 변형을 통해 BAC과 같은 유사한 특성을 가진 P1 유래의 인공 염색체 (PAC) 가 새로운 벡터 시스템으로 개발되었다(Ioannou *et al.*, 1944). 따라서 현재 유전체 연구를 위한 벡터 시스템은 BAC과 PAC이 주로 이용되고 있다.

인간 유전체 및 다른 생물 유전체의 고밀도 물리지도를 작성하기 위한 클론의 순서화(Assembly) 작업엔 다양한 기술이 응용되고 있다. 제한효소 처리에 의해서 나타나는 단편들의 유사성 분석이 *C. elegans*, 람다 클론, *S. cerevisiae* 및 *A. thaliana*의 순서화 작업에 응용되었다(Coulson *et al.*, 1988, 1991; Olson *et al.*, 1986; Riles *et al.*, 1993; Hauge & Goodman, 1992). Cross hybridization 방법 또한 클론의 순서화 작업에 응용되고 있다(Ward& Jen, 1992). 특히 이

방법은 YAC 클론과 코스미드 클론 사이의 틈을 메우는 데 중요한 기술로 평가되고 있다. 또한, 작은 틈의 연결은 long PCR이라는 기법을 응용하는 것이 그 대안적인 방법이 될 것이다. 물리지도 작업은 100-1000 bp를 특이적으로 증폭할 수 있는 PCR 기법이 도입되면서 한층 힘을 얻게 되었다. 보통 이와 같이 증폭되는 단편을 서열 꼬리표 부위(Sequence Tag Site, STS)라고 일컫는다(Olson *et al.*, 1989). 이들은 유전체 한 개의 특정 부위에서 발생하는 것으로 YAC과 BAC의 정렬화 작업에 응용되고 있다. STS 서열은 임의 선별 유전체 클론과 cDNA 클론을 분석함으로써 또는 다형현상을 보이는 유전적 마커를 분석함으로써 얻을 수 있다. 이와 같은 노력으로 인간 유전체를 cover하는 약 3만개로 구성된 STS 지도가 보고 되었다(Hudson *et al.*, 1995; Schuler *et al.*, 1996). 이것은 STS간 100 kb의 평균 거리를 갖는 것으로 인간 유전체 계획을 위한 중요한 실마리를 제공하였다(Collins & Galas, 1993). STS 지도는 또한 고품질 BAC 유전자 은행의 탐색을 위한 특정 유전체 부위의 클로닝을 가능하게 해주었다.

결국, 물리지도는 전체 유전체 해석 계획의 근간일 뿐만 아니라 특정 유전체 부위에 대한 주소와 같은 역할을 하여 유전자 클로닝을 보다 쉽고 용이하게 해주는 결과물이다. 이것은 궁극적으로 어떤 질병과 관련된 유전자의 유전적 위치를 제공하여 직접적으로 클로닝을 편리하게 해주는 자료이기도 하다.

2) 임의 선별을 이용한 간접 유전체 해석법

큰 유전자 단편의 염기서열 결정하는 데에는 두 가지의 염기서열 분석 전략이 있다. 이것은 무작위 방식(Random method)과 직접 방식(Direct method)이다. 두 방법은 클로닝 방법, insert 크기 및 염기서열 전략상에 차이가 있다. 무작위 방법은 M13 파지나 플라스미드에 의해서 클로닝된 시료를 사용한다. 대량의 이들 클론들은 표준 벡터 특이적인 primer에 의해서 염기서열을 결정하여 원래의 단편(BAC 또는 BAC insert)에 대한 무작위적인 염기서열 결과를 도출한다. 따라서 이와 같은 결과가 연속성을 갖기 위해서는 정열화 작업을 반드시 수행해야만 한다. 무작위 염기서열 결정법은 목적 단편의 75-90%까지 수행되며 이후 틈새 연결을 위한 직접 법 전략이 후속 방법으로 사용된다.

큰 크기의 유전체 절편을 단편화 작업(Fractionation)을 통해서 서브 클로닝 하는 방법은 제한효소 절단법(Messing *et al.*, 1981), DNase I 처리법(Anderson *et al.*, 1981), 초음파 분쇄법(Deininger, 1983), HPLC(Oefner, *et al.*, 1996) 또는 Nebulisation 등 다양한 방법이 소개되고 있다. 제한효소 법은 사용하기 쉬운 반면에 몇 가지 단점이 있다. 예를 들면 한 종류의 효소에 의한 완전 절단은 non overlapping 클론을 생산한다는 것이다. 따라서 이를 극복하기 위해서 부분 절단 법이나 두 가지 이상의 효소를 이용하는 것이 그 대안으로 쓰여지고 있다. 그러나 이것은 목적 유전체내에 효소 인식 부위가 적다면 완전한 유전체 작성이 어렵다는 또 다른 난관에 빠진다.

산탄 유전자 염기서열 결정법은 대규모 염기서열 결정 계획에 사용된 고전적인 방법이다. 이 방법으로 인간 미토콘드리아 유전체(Anderson *et al.*, 1982)와 람다 파지 유전체(Sanger *et al.*, 1982) 및 아데노 바이러스 유전체(Gingeras *et al.*, 1981)가 해석되었다. 이후로 무작위 산탄 염기서열 분석법은 다양한 생물의 유전체 계획에 응용되고 있다. 전형적인 무작위 산탄 염기서열 결정법은 관심이 있는 클론을 포함하는 많은 수의 클론을 대상으로 하는 것이다. 대체로 산탄 유전체 염기서열 결정을 위한 coverage는 코스미드의 경우에는 6-8배 (redundancy)로 수행된다. 따라서 요구되는 염기서열 반응 숫자가 많은 경우엔 직접 전략법 보다 훨씬 더 유용한 전략으로 평가되고 있다. 또한 이 방법은 똑같은 primer에 의해서 표준화된 방법으로 수행 될 수 있는 장점이 있다. 그러나 산탄 유전체 서열 분석법은 최적화된 컴퓨터 알고리듬에 기초로 하기 때문에 이들을 지원 할 수 있는 컴퓨터 시스템이 구비되어야 하는 부담이 있다. 산탄법에 의해서 나오는 결과를 기초로 한 2차 방법인 틈새 메우기 방법은 대게 서브 클로닝이나 PCR 방법에 의해서 최종적으로 마무리 된다(Wilson *et al.*, 1992).

3) 직접 유전체 해석법

직접 법에 의한 염기서열의 기술적 사용은 목적 유전자가 위치한 유전체 부위에 대한 정보를 알고 있을 때 가

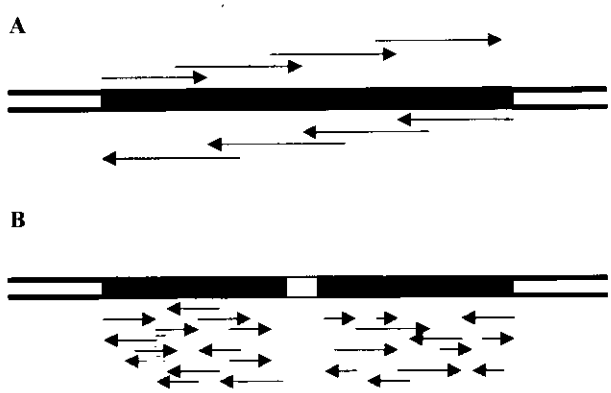

Fig. 2. The principal differences between primer walking and shotgun sequencing. In primer walking (A), sequence data from each reaction are used to design new sequencing primers. A minimal redundancy and full coverage is obtained. in shotgun sequencing (B), randomly selected and assembled to form a contigous sequence. Gaps might have to be filled by directed sequencing methods.

능하다. 이와 같은 전략은 산탄 유전체 분석법과 비교할 때 최소한의 염기서열 반응만이 요구된다는 장점을 갖는다. 따라서 보다 정교한 클로닝 과정과 세심한 목적 유전자의 선별 작업이 선행적으로 수행되어야 한다(Fig. 2).

직접 유전체 해석을 위한 가장 일반적인 접근법은 primer walking 법이다. 이 전략은 한번의 반응에 의해서 나온 염기서열 결과를 기초로 다음 단계의 반응에 필요한 primer를 제작하고 이를 이용해서 다음의 염기서열 반응을 연속적, 반복적으로 하는 과정을 일컫는 것이다(Strauss *et al.*, 1986). 또한, 이와 같은 방법은 양쪽 방향 모두에서 진행하는 것이 가능하여 그 결과의 정확성을 높일 수 있으며 각 반응에서 얻어진 염기서열 결과를 알고 있으므로 산탄법과 비교할 때 정렬화 작업이 용이한 장점을 갖는다. 반면에 단점으로는 연속적으로 새로운 primer의 작성이 요구된다는 것이다. primer의 합성은 값이 비싸면서 느리다는 것이며 합성된 primer는 한 종류의 목적 벡터에만 사용 가능하다는 것이다. 이와 같은 단점을 극복하기 위한 방법으로는 짧은 primer 조합을 갖는 primer 은행을 사용하는 방법이다(Studier, 1989; Szybalski, 1990). 그러나 이 방법은 이론적으로 shortmer들이 염기서열 분석에 사용 가능하다는 연구 결과가 있음에도 불구하고 많은 연구자들에게 설득력 있게 이용되고 있지 않다. 또 다른 대안으로는 nested deletion 방법이 있다(Henikoff, 1984; Misra, 1985). 이것은 한가지의 primer를 사용할 수 있는 장점이 있으나 실험 과정이 까다롭다는 단점을 가지고 있어 유전체 해석에는 많이 사용하지 않고 있다.

직접 분석 전략은 대체로 클론 사이의 틈을 메우는 작업, 즉 산탄 분석법을 수행한 후 최종 단계에 주로 사용된다. 따라서 대규모 염기서열 분석 계획에서 직접 분석법은 다음과 같은 이유로 드물게 이용된다. 첫째, 직접 분석법은 정확한 물리지도가 준비되어야 하며 목적 부위의 정확한 클로닝이 수반되어야 한다. 둘째, primer 합성이 상업화 되면서 그 비용이 많이 낮아 졌음에도 불구하고 경제적 부담이 크다는 것이다. 마지막으로 직접 전략은 산탄 염기서열 분석법에 비해서 숙련된 연구원과 표준화된 방법이 요구된다는 것이다. 그럼에도 불구하고 직접 분석법에 의한 많은 논문들이 보고되고 있다. *S. cecevisiae* 유전체 분석 계획의 일부를 담당하는 유럽 협의체는 2.8배 (Wimann *et al.*, 1993)와 2.6배 (Voss *et al.*, 1995)의 redundancy를 갖는 두개의 코스미드를 primer walking에 의해서 해독하였다. 또한, 0.8 Mb의 유전체를 갖는 *M. pneumoniae*가 primer walking, nested deletion 및 산탄 분석법에 의해서 해독되었다(Hilbert *et al.*, 1996). 이 계획은 최종 2.95배의 redundancy 분량이었으며 5,095개의 염기서열 결정용 primer가 사용되었다. 또한, 지속적인 연구를 통해서 산탄 분석법과 직접 분석법을 결합시킨 순서화된 산탄 분석법 (Ordered shotgun sequncing, OSS)이 개발되었다(Chen *et al.*, 1993). 이것은 지속적인 연결작업과 각 클론들의 end sequence 결과에서 확보된 클론들의 중첩 부위를 목적 유전자로 선별한 후 산탄법으로 전체 목적 유전체의 염기서열을 결정하는 방법이다.

4) 어떻게 유전체 염기서열을 결정할 것인가?

전체 유전체 분석에 적절한 방법이 무엇이냐 하는 질문에 대해서는 아직까지 명쾌한 답이 없다. 그러나 확실하건데 물리지도 작성이 가능하다는 가정이 성립된다면 가장 선호하는 방법은 직접 분석법이 될 것이다. 많은 염기서열 분석법과 새로운 분석 기술의 개발은 *S. cerevisiae, C. elegans*와 같은 전체 유전체 염기서열 해독이라는 결과물을 도출하였다(Thierry *et al.*, 1995; Johnston, 1996; Sulston *et al.*, 1992; Wilson *et al.*, 1994). 이것은 현재까지 확립된 모든 가능한 방법에 의해서 완성된 것이다.

전체 유전체의 산탄 염기서열 분석법은 Venter가 책임자로 있었던 TIGR(The Institute of Genome Research)라는 생명공학 벤처회사에서 처음으로 응용하였다. 1995년에 1.83 Mb의 유전체를 갖는 *H. influenzae*의 전체 염기서열이 해독되었다(Fleischmann *et al.*, 1995). 이것은 1.6-2.0 kb의 insert를 갖는 플라스미드 24,000개를 무작위 선별 염기서열 분석법과 람다 클론, PCR 산물의 직접 방법에 의해서 완성된 결과이다. 전체적인 redundancy는 6.3배이었으며 그 정확도는 1,000-5,000 bp당 한 개의 오류를 보일 정도로 매우 정확하였다. 또한, 매우 유사한 염기서열 전략으로 1.66 Mb의 유전체를 갖는 *M. jannaschii*의 전체 염기서열이 결정되었다(Bult *et al.*, 1996). 특히 이 유전체 계획의 특징은 16 kb의 insert를 갖는 람다 클론의 end-sequencing 방법과 플라스미드의 직접 염기서열 결정법이 동시에 진행되어 완성되었다는 것이다. 현재 박테리아 및 원생 박테리아 몇 종에 대한 전체 유전체 산탄 분석법이 진행 중에 있다(Table 2).

인간 유전체의 염기서열 분석 작업은 상기에서 논의된 모델 생명체로부터 많은 경험을 축적하면서 진행되었다. 이와 같은 노력으로 물리지도를 기초로 한 clone by clone 이라는 방식을 적용하여 *S. cerevisiae*와 *C. elegans*의 유전체가 해독된 것이다. 그러나 연구자들에게 이용 가능한 물리지도는 YAC을 근간으로 한 것이어서 충분한 해상도를 주지 못하였다(Boguski *et al.*, 1996). 따라서 BAC, PAC 및 P1 클론을 기초한 high-redundant sequence-ready map이 완전한 유전체 해독 작업에 절실이 필요하게 되었다. 1997년 전체 유전체 산탄 방법이 인간 유전체 계획에 제안되었다. 무작위로 선별된 인간 DNA의 크고 작은 플라스미드의 조합을 전체 유전체의 10배의 coverage로 작성

한 후 염기서열 분석이 진행되어야 한다는 것이며 하나의 생물정보학 그룹에 의해서 정렬화 작업이 수행되어야 한다는 주장이었던 것이다(Weber & Myers, 1997). 이와 같은 전략의 장점은 물리지도 작성을 하지 않아도 된다는 것과 다른 벡터로 인간 DNA를 클로닝하는 것보다 플라스미드로 클로닝이 쉽다는 일반적인 사실에 기초로 한 것이다. 근본적으로 엄청난 비용을 줄일 수 있는 방법이기도 하였다. 그러나 이와 같은 제안은 산탄 분석 과정이 매우 까다롭다는 논쟁을 불러 일으켰다(Green, 1997). 다른 측면으로 위에서 설명한 clone by clone 전략은 다음과 같은 장점을 제공한다. 첫째 다른 연구 그룹에 클론들을 분양하기가 용이하다는 점 둘째, 제한효소에 의해서 단편들의 크기를 쉽게 조절할 수 있다는 점 등이다. 또한 큰 크기의 insert를 갖는 클론이 효율적으로 틈새 메꾸기 전략에 용이하다는 가장 큰 잇점을 제공한다. 전체 염기서열을 산탄법으로 수행할 때에는 질병 유전자와 같은 관심 있는 유전자의 위치 클로닝(Positional cloning)이 다른 방법에 비해서 어렵다는 것이며 많은 클론들의 유지가 힘들다는 것이다. 그럼에도 불구하고 1997년 Venter는 인간 유전체 계획 초기에 산탄 방법을 적용하여 3년 후인 금년에 인간 유전체의 전체 염기서열 초안을 발표하였다. 그러나 이와 같은 산탄법이 성공하려면 유전체의 큰 부분을 연결하는 contig 구성체인 BAC과 BAC end-sequencing 기술이 확립되어야 한다.

4. cDNA 염기서열 결정법

전체 유전체 분석의 궁극적인 목표는 유전체 안에 존재하는 유전자 정보를 해독하는 것이다. 그러나 이것은 각기 다른 유전체에 존재하는 유용한 정보의 변이가 심한데 그 어려움이 따른다. 박테리아 유전체는 작으며 전체 염기서열 방법 중 산탄법에 의해서 효과적으로 해석될 수 있다. 그러나 인간 유전체에서 단백질을 암호화하는 부분은 전체의 3% 정도이다. 이것은 상당히 낮은 비율이며 그만큼 기능을 갖는 유전자의 해석이 용이하지 않다는 것을 의미하는 것이다. 진핵생물의 유전자 동정은 이와 같은 이유로 해서 상당히 어려운 작업이다. 또한 진핵생물에 존재하는 intron이라는 서열로 인해서 유전자의 해석이 더욱 더 어렵게 된다. 따라서 진핵생물의 유전체 안에 존재하는 유전자를 동정하는 새로운 기법인 ESTs 염기서열 분석법이 대안적인 방법으로 대두되었다. ESTs는 cDNA의 부분 염기서열을 일컫는 용어로 mRNA를 역전사하여 intron이 없는 유전자의 암호화 부분만을 제공하여 직접적으로 유전자 동정에 긴요한 방법으로 인정되고 있다. 최근 몇 년간 대규모 ESTs 염기서열 분석 계획이 수행되어 유전체 분석에 지대한 영향을 끼쳤으며 유전자의 동정 뿐만 아니라 물리지도 작성과 발현 profiling이라는 새로운 영역이 개척되는 계기를 마련하였다. 인간 유전체 계획의 주요 목적은 모든 인간 유전자의 완전한 동정이다. 이것은 다른 유전체 계획에도 마찬가지 일 것이다. ESTs 염기서열 분석법은 동정된 유전자 수가 증가하면서 그리고 인간의 다양한 조직에서 많은 계획들이 도출되면서 호평받는 기술이 되었다. 결국 많은 결과들이 보고 되면서 dbEST라는 유전자 정보 은행이 설치되었으며 (Boguski *et al.*, 1993) 현재 약 2 million entry가 등록되었다. 또한, 여기엔 인간의 다양한 조직에서 밝혀진 신규 유전자가 60% 나 포함되어 있다.

1) mRNA

원핵생물과 진핵생물의 유전자 발현 패턴에는 몇 가지 차이가 존재한다. 원핵생물의 유전자는 오페론이라는 단위로 유사한 기능을 갖는 유전자들이 정렬되어 있어서 전사 및 발현이 동시에 이루어진다. 그러나 진핵생물의 유전자는 서로 다른 염색체에 위치하며 각각의 유전자의 전사 및 발현이 대체로 개별적으로 이루어진다. 특히 전사 단계에서는 인트론 부위가 스플라이싱 이라는 기작을 통해서 제거된 후 성숙한 mRNA 전사체가 형성된다. 이 과정에서 CAP(7-methyl-GTP)이라고 하는 구조가 전사체의 5' 부위에 붙게 되며(Shatkin, 1976) poly(A)-tail이라는 구조가 3' 부위에 붙게 된다(Lim & Canellakis, 1970). 성숙한 mRNA는 그리고 나서 세포질로 이동하여 단백질로 번역된다. 또한, 유전자를 암호화하는 서열은 UTR이라는 번역되지 않는 서열 사이에 위치하게 되는 특징 또한 원핵생물의 유전자와는 다른 특징이다. 진핵생물의 poly(A)-tail의 존재는 행운이었다. 왜냐하면 그것은 쉽게 mRNA를 분리 가능하게 해주었기 때문이다. 결과적으로 이것은 특정 조직에서 특정시기에 발현하는 모든 유전자들을 대표하는 유전자 은행의 확보가 가능하게 된 계기를 주었던 것이다.

2) cDNA 유전자 은행의 구축

cDNA 분석의 성공 여부는 클로닝 방법과 목적 여부에 달려있다. 여러 종류의 클로닝 방법이 ESTs 염기서열 분석과 발현 패턴 분석 및 전체 cDNA 클론의 확보를 위해 응용되고 있다. 그러나 일반적으로 ESTs 분석의 성공은 시작 초기에 구축한 유전자 은행이 똑같은 빈도로 mRNA 집단에 존재하는 모든 서열을 대표하는 지의 여부에 달려있다.

mRNA의 분리는 일반적으로 전체 RNA를 분리하는 단계에서부터 시작된다. 진핵세포에서 전체 세포내 RNA의 1-5% 만이 mRNA이며 대다수는 ribosoml RNA 및 transfer RNA가 차지한다. 모든 RNA 종들은 매우 민감하며 쉽게 RNase에 의해서 분해된다. 따라서 실험에 쓰이는 기구나 시약은 반드시 RNase에 저항성을 갖는 시약이나 효소를

처리하여 사용해야 한다. 시료 세포를 모으는 단계에서 Rapid-snap-freezing이라는 방법은 RNA 분해를 효과적으로 막는 데 하나의 대안이 될 수 있다. RNA는 일반적으로 GTC(guanidium salt)로 처리 한 후에 페놀 및 클로로포름 추출에 의해서 분리 할 수 있다(Chirgwin *et al.*, 1979; Chomczynski & Sacchi, 1987). 대부분의 성숙한 mRNA는 약 20-200개의 연속적인 아데닌을 갖는다. 이런 특징을 이용하여 poly(A)에 상보적인 oligo(dT) probe를 cellulose (Aviv & Leder, 1972), latex particle(Hara *et al.*, 1991) 및 magnetic bead(Hornes & Korsnes, 1990; Jakobson *et al.*, 1990)와 같은 solid phase에 결합시켜 쉽게 mRNA를 분리할 수 있게 되었다.

mRNA를 dsDNA로 전환하여 클로닝 하는 방법이 소개되고 있다. mRNA는 여기에 상보적인 first strand에 poly(A) tail이 결합할 수 있는 oligo(dT)-primer 및 특정 유전자에 상보적인 유전자 특이적인 primer(Frohman *et al.*, 1988), random hexamer(Dudley *et al.*, 1978) 등을 붙혀 역전사 효소에 의해 ssDNA로 전환될 수 있다. 이후 second strand DNA의 합성을 위해서 개발된 첫 번째 방법은 second strand의 연속적인 신장에 의해서 형성된 일시적인 hairpin-loop를 이용하는 것이었다(Efstratiadis *et al.*, 1976). 이 loop 구조는 S1 nuclease에 의해서 제거되며 ds cDNA의 끝 부분은 T4 DNA polymerase에 의해서 클로닝 단계 바로 전에 repair 할 수 있다는 방법이다. 현재 자주 이용되는 또 다른 방법은 몇 가지의 효소를 조합하는 것이다 ((1) RNA-DNA 혼성체의 RNA 부분을 RNase H로 분해하는 것 (2) DNA polymerase I을 사용하여 신장을 위한 primer로서 nicked RNA를 사용하는 것 (3) 새로 합성된 DNA에 존재하는 틈을 DNA ligase를 사용하여 연결하는 것).

cDNA는 일반적으로 플라스미드나 람다 파지로 클로닝 된다. 파지는 넓은 범위의 insert DNA를 클로닝 할 수 있어서 상대적으로 품질이 좋은 유전자 은행과 full lengh cDNA 전사체를 클로닝 하는데 유용하게 사용된다. 또한, 구축된 유전자 은행의 스크리닝 단계에서도 박테리아 콜로니 보다 파지 플라크로 하는 것이 보다 손쉬운 방법이다. 그러나 플라스미드는 주형 DNA의 준비 및 염기서열 결정 반응 등에서 손쉬운 조작이 가능한 장점이 있다. 따라서 플라스미드로 전환이 가능하게 조작된 즉, *in vivo* excision이 가능한 파지 벡터가 현 EST sequencing 계획에 주로 사용되고 있다(Short *et al.*, 1988). cDNA 분석에 기초를 둔 많은 응용분야에서 가장 중요한 것들 중 하나는 insert의 방향성을 알 수 있는가 하는 것이다. 따라서 방향성 갖게 클로닝 하는 방법이 first strand 합성 단계에서 oligo(dT)-primer 부위에 *Not*I 제한효소를 삽입함으로써 가능하게 되었다. 이후 second strand 합성시에 *Eco*RI overhang을 갖는 adaptor를 cDNA 양쪽 끝에 연결하여 벡터에 클로닝 하면 첫 단계에 사용된 메칠화된 nucleotide가 안쪽에 위치한 제한효소 부위를 효과적으로 blocking 하므로 최종적으로 합성된 cDNA 클론들의 방향성을 부여하게 되는 것이다. PCR 기법 또한 cDNA 집단의 증폭을 위해서 이용되고 있다(Akowitz & Manuelidis, 1989). 그러나 불행하게도 이 방법은 원래의 전사체에 대해 변형된 밀도를 나타내는 오류가 발생할 수 있다는 단점이 있다.

cDNA 유전자 은행의 품질은 다른 방법으로 평가할 수 있다. 클로닝 하기 전에 mRNA 집단의 품질은 일반적으로 모든 세포에 존재하며 그 크기 또한 잘 알려진 actin이라는 유전자를 probe로 사용하여 northern blotting에 의해서 결정할 수 있다. Actin 유전자에 상응하는 RNA가 분해되지 않고 많은 양의 밴드로서 나타나고 전체 mRNA가 0.4에서 4 kb 사이에 밀집되어 존재한다면 구축된 cDNA 유전자 은행의 품질은 좋다고 평가할 수 있다(Moreno-Palanques & Fuldner, 1994). 또한, cDNA 유전자 은행의 비 편중성은 발현 수준이 다른 알려진 유전자를 probe로 이용하여 스크리닝을 통해서 평가 할 수 있다(Okubo *et al.*, 1992). 즉, 양성 플라크의 출현 비율은 northern 실험에서 처럼 시그날의 감도가 같아야 한다. 또한 좋은 품질의 유전자 은행은 chimeric 클론의 비율이 최소화 된 것이어야 한다. 이것들은 클로닝 과정시에 사용한 벡터를 최소한의 양으로 사용하거나 다량의 adaptor를 사용함으로써 해결 할 수 있다. 또한, cDNA 유전자 은행 내에 존재하는 스플라이싱 되지 않은 미 성숙한 mRNA 전사체들은 milder extraction protocol (NP40 detergent)를 사용함으로써 최소화 할 수 있다(Assheim *et al.*, 1994).

3) 단백질을 암호화하는 유전자의 동정

cDNA 유전자 은행의 구축 전략과 여기에서 생산된 클론의 염기서열 결정 전략은 cDNA 분석 결과에 직접적으로 영향을 주는 중요한 요인이 된다. 초기 ESTs 염기서열에 대한 부단한 연구로부터 random primed cDNA 유전자 은행이 유전자 동정을 위한 최적의 방법이라는 것이 여러 연구 논문들로부터 확인되었다. Poly(A)-tail을 이용한 염기서열 결정의 문제가 이 방법으로 최소화 할 수 있었으며 이로부터 나온 결과는 3' 또는 5' UTR 보다 오히려 단백질암호화 부위가 대다수를 차지하였다. 따라서 펩타이드 상동성 분석을 이용한 기능 분류가 쉽게 되었다(Hoog, 1991; McCombie *et al.*, 1992; Waterston *et al.*, 1992). 또한, 몇몇 연구자들의 부단한 노력으로 방향성을 갖는 새로운 cDNA 유전자 구축법이 탄생하게 되었다. 이 방법으로 cDNA의 5' 끝 부분부터 염기서열 분석이 가능하게 되었다. 유전자의 동정은 대부분의 클론이 5' 부위에서 truncation 되기 때문에 상대적으로 많은 수를 반복 수

행해야 하며 동시에 무작위 선별이 최적화되어야 한다. 무작위 염기서열 결정의 확실한 장점은 선별된 클론을 단순하게 작업할 수 있다는 것이다. 3' end로 부터의 염기서열 분석기술에 대한 많은 논문들이 보고되고 있다(Hofte *et al.*, 1993). 그러나 3' end로 부터 해석된 염기서열 결과는 새로운 유전자의 동정과 상동성 분석 및 유전자 기능 분석에 적절하지 않은 것으로 평가된다. 더구나 앞서 언급한 것처럼 polymerase silppage 현상으로 poly(A)-tail을 통한 염기서열 분석의 어려움 때문에 성공율 또한 상당히 낮은 단점이 있다.

ESTs는 물리지도 작성에도 이용 될 수 있다(Wilcox *et al.*, 1991). 모든 ESTs는 intron이 배제되어 있으며 독특한 염기서열을 제공하므로써 STS로서 이용 가능한 잠재적인 클론이 된다. Intron은 3' UTR에는 드물게 존재하므로 3' end로 부터 해석된 염기서열이 물리지도 작성에 아주 적당하다(Matsubara & Okubo, 1993; Lanfranchi *et al.*, 1996). 또한, ESTs는 somatic hybrid 연구의 촉매제가 될 수 있으며 유전자 family 사이의 특징을 구별 가능하게 해준다. Poly(A)-tail을 이용한 염기서열 결정 방법의 대안으로써 3' 부위의 클로닝을 위한 새로운 방법이 개발되었다. 이 방법은 효과적으로 3' 염기서열을 발생시키므로 물리지도 작성 뿐만 아니라 유전자 발현 profiling에 이용된다.

전형적인 체세포의 mRNA는 발현 양에 따라서 3가지 종류로 분류할 수 있다. 가장 높은 발현을 보이는 10-15 종류의 유전자들은 전체 mRNA의 10-20%를 차지하며 1,000에서 2,000개의 유전자들은 중간적인 발현 수준을 보이며 전체 mRNA의 40-45%를 차지한다. 15,000에서 20,000개의 달하는 유전자는 가장 적은 발현 양을 갖는 것으로써 전체 mRNA의 40-45%를 차지한다. 따라서 ESTs 분석에 있어서 cDNA 클론의 무작위적인 선별은 가장 발현 빈도가 높은 유전자들이 선별되는 경향을 보인다. 그러므로 이와 같은 cDNA의 반복성을 줄이기 위한 한 가지 방법은 전사체의 양이 균일한 cDNA 유전자 은행을 구축하는 것이다(Ko, 1990; Patanjali *et al.*, 1991; Soares *et al.*, 1994).

4) 유전자의 발현 분석

특정 유전자 은행을 이용한 대량의 ESTs 염기배열 해독 작업은 특정 조직에서 연구된 유전자 발현의 복잡성과 발현 수준에 직접적으로 상응하는 하나의 유전자 집단을 대상으로 진행되어야 한다. 똑같은 유전자로부터 유래된 염기서열은 하나의 cluster로 grouping될 수 있다. 따라서 하나의 그룹 내에 존재하는 클론의 수는 목적 집단에서 그 유전자의 발현 빈도를 대별하는 것이다(Rounsley *et al.*, 1996). Random primed 유전자 은행은 실제 발현 수준을 대별하지 못한다. 왜냐하면 5' end가 종종 cDNA 유전자 은행에서 truncation되는 경우가 많기 때문이며 5' 염기서열은 다른 starting point를 갖는 똑 같은 전사체로 부터 나오는 예가 많기 때문이다. 따라서 5' 염기서열은 잘못된 발현수준을 보이는 경우가 많다. 그렇다고 아직까지 유용한 방법이 있는 것은 아니다. 대안적으로 조직들 간의 metabolic 차이점은 기능적 catalogue로 유전자들을 groupping함으로써 얻을 수 있다(Adams *et al.*, 1993; Liew *et al.*, 1994).

최적의 발현 profiling은 각각의 유전자가 정확하게 같은 부위에서 유래된 염기서열 일 때에 얻어질 수 있다. 발현 연구를 위한 대규모 ESTs 염기서열 분석은 상당히 더딘 과정이며 비용인 많이 드는 단점이 있다. 1995년에 소위 유전자 발현의 연속적 분석(Serial analysis of gene expression, SAGE)이라고 하는 방법이 소개되면서부터 이와 같은 단점이 일부 극복될 수 있었다. Type II 제한효소를 사용하는 이 방법의 전략은 cDNA의 짧은 꼬리표 연결체(Concatemer)를 클로닝하고 염기서열을 분석하는 것이다(Velculescu *et al.*, 1995). 9 mer 정도가 되는 염기서열 꼬리표(Sequence tag)는 인간 유전체의 95% 이상을 동정하기에 충분한 것이다. 일차적으로 각각의 염기서열이 이와 같은 방법에 의해서 30-40개의 꼬리표를 생산하며 이들을 이용하여 각각의 유전자를 동정한다. 다음으로 정상 및 비정상 세포에서 발현되는 꼬리표를 분석하는 것이다. 또 다른 방법은 3' end로 부터 얻어진 염기서열 결과를 최적화한 pyrosequencing이라는 것이다(Ronaghi *et al.*, 1998). 이 방법은 대규모 계획이 가능한 완전하게 새로운 non-gel based 염기서열 분석법이었으나 아직까지 상업적으로 가능하지 않다.

상기에서 기술된 ESTs 및 sequence tag 염기서열 방법은 완전한 유전자 발현 profile을 생산한다. 또한 목적 유전자가 다른 조직에 존재하는지 아닌지 까지 분석 가능하게 해 줄 뿐만 아니라 유전자의 up/down regulation까지 분석하는 것이 가능하다. 이것은 다른 발현 분석을 위한 기술과 비교할 때 상당한 장점이다. 또한, 완전한 발현 분석을 위한 방법이 소개되고 있다. 일명 cDNA microarray라는 기법이다(Schena *et al.*, 1995; Drmamac *et al.*, 1996). Yeast의 6,000개의 유전자 대부분을 PCR을 통해서 증폭한 다음 하나의 유리판 위에 배열한 후 다른 증식 조건으로부터 yeast RNA를 준비하고 혼성화 반응을 통해서 정량적인 발현 profile이 분석 가능해 진 것이다(DeRisi *et al.*, 1997).

시간, 비용 및 염기서열 분석 능력이 제한을 받을 때에는 소위 differential display (Liang & Pardee, 1992), RNA fingerprinting (Welsh *et al.*, 1992), arbitrarily difference anlaysis (RDA) (Hubank & Brown, 1991), substraction

hybridization (Wang & Brown, 1995) 방법들을 사용 할 수 있다.

5) Full length 염기서열 분석법

전체 단백질을 암호화하는 염기서열의 결정은 유전자의 기능적 특성을 위해서 중요한 단계이다. Full-length cDNA 염기서열 전략은 목적 유전자의 자세한 분석이 선행되어야 한다. 부분 염기서열을 알고 있다면 초기에는 선별된 cDNA 염기서열이 유래된 클론의 완전한 insert를 재분석하는 노력이 필요하다. 전통적으로 plaque hybridization 방법이 cDNA 유전자 은행으로부터 보다 긴 단편을 분리하는데 이용되어 왔으나 이것은 시간이 많이 요구 된다는 단점이 있다. 비록 많은 full-length cDNA를 클로닝 할 수 있는 방법이 개발되었음에도 불구하고 대부분의 전사체는 불완전한 first strand 합성으로 인해서 5' end 부위가 truncation 되어 있다.

특별히 5' 및 3'의 확보하지 못한 부위를 결정하기 위해서 PCR 기법을 근간으로 하는 다양한 기법이 소개되었다. cDNA end의 신속한 증폭이라고 하는 방법(Rapid Amplification of cDNA End, RACE)이 처음에 개발된 것이다(Frohma *et al.*, 1988). 이후 거의 유사한 anchored PCR 기법이 소개되었다(Loh *et al.*, 1989). 확보하지 못한 3' end를 찾기 위해서 first strand 합성 과정에서 5' end에 PCR primer 부위를 갖는 oligo(dT)로 PCR 반응을 수행한 후 oligo(dT)에 결합된 primer와 부분 염기서열 결과로부터 합성된 유전자 특이적인 primer(Gene Specific Primer, GSP)에 의해서 후보 유전자들이 증폭된다. 이와 같은 방법을 3' RACE라고 한다. 5' end를 확보하기 위해서 GSP가 first strnad 합성과정에 이용된다. 5' end 부위에 primer 부위가 도입되고 GSP와 새로 삽입한 primer를 이용하여 PCR을 수행하면 5' end를 클로닝 할 수 있다. 이때에 일반적으로 사용되는 새로운 primer 부위는 terminal transferase에 의한 homopolymer가 삽입되며 이러한 방법을 5' RACE라고 한다.

Full-length 염기서열 분석의 속도는 EST 및 차별화 발현 기술에 의해서 생산된 새로운 유전자의 클로닝 속도와는 비교되지 않을 정도로 빠르다. 이것은 상기에서 기술한 것처럼 PCR이라는 강력한 기법을 응용할 수 있기 때문에 가능해 진 것이다. 그러나 아직까지 full-length 및 고품질 염기서열은 잠재적으로 고도로 정확한 database의 비교 분석과 유전자의 기능과 구조를 예측할 수 있는 프로그램에 의해서 주로 확보되고 있다.

결 론

이제 21세기는 기능 유전체학(functional genomics)이라는 새로운 학문이 생명과학 관련 전분야에 걸쳐 그 주류를 담당할 것으로 많은 연구자들은 예측하고 있다. 이것은 이미 종료된 그리고 진행중인 다양한 생물의 전체 유전체 해석이 가능해 졌기 때문이다. 지금까지는 한 개의 유전자에 대한 집중적 연구가 진행되어 왔으나 앞으로는 다양한 유전자의 상호 작용에 관한 연구로 그 흐름이 바뀌 질 것이다. 따라서 앞으로의 유전자 및 유전체 연구는 다각적인 시각에서 접근되어야 할 것이다. 대략적으로 인간의 1/6 정도인 500 Mb의 크기를 갖는 것으로 알려져 있는 누에의 유전체를 해석한다면 상기에서 기술된 것처럼 많은 비용과 인적자원 그리고 미래 지향적인 컴퓨터 운영체계(Bioinformatics, 흔히 생물정보학이라고 한다) 등이 완벽히 구비되어야 할 것이다. 따라서 현재의 여건에서는 전체 유전체를 해독하는데 집중하는 것보다는 누에의 기능 유전자 대량 발굴 계획이 보다 설득력이 있을 것으로 본다. 현재의 세계적 흐름은 유전자 전쟁이라고 불릴 정도로 어떻게, 어떤 유전자를 찾아서 그 기능이 무엇인지를 먼저 밝히는데 치열한 경쟁을 하고 있는 것이다. 이러한 측면에서 피브로인, 세리신등 누에 특유의 유전자 혹은 산업화 가능성 후보 유전자들을 가능한 많이 동정할 수 있는 대규모 ESTs project, SEGA, 혹은 full-length cDNA project가 현 시점에서 가장 효과적인 것이라 생각된다.

참고문헌

Aasheim, H. C., Deggerdal, A., Smeland, E. B., Hornes, E., 1994. A smiple substraction method for the isolation of cell-dpecific genes using magnetic monodisperse polymer particles, *Biotechniques* **16** : 716-721.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., 1991. Complementary DNA sequencing: expressed sequencing tags and human genome project. *Science* **252** : 1651-1656.

Adams, M. D., Kerlavage, A. R., Fields, C., Venter, J. C., 1993. 3400 new expressed sequence tags identify divesity of transcripts in human brain. *Nature* **355** : 632-634.

Akowitz, A., Manuelidis, L., 1989. A novel cDNA/PCR strategy for efficient cloning of small amounts of undefined RNA. *Gene* **81** : 295-306.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., Young, I. G., 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290** : 457-140.

Anderson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M., Naslund, A. K., Eriksson, A. S., Winkler, H. H., Kurland, C. G., 1998. The genome sequence of *Rikettsia prowazeki* and the origin of mitochondria, *Nature* **396** : 133-140.

Ansorge, W., Sproat, B. S., Stegemann, J., Schwager, C., 1986. A non-radioactive automated method for DNA sequence determination. *J. Biochem. Biophys. Methods* **13** : 315-323.

Aviv, H. and Leder, P., 1972. Purification of biologically active globin messenger RNA by chromatography on oligothymidly acid-cellulose. *Proc. Natl. Acad. Sci. USA* **69** : 1408-1412.

Boguski, M. S., Lowe, T. M., Tolstoshev, C. M., 1993. dbEST database for 'expressed sequence tags'. *Nat. Genet.* **4** : 332-333.

Boguski M., Chakravarti, A., Gibbs, R., Green, E., Myers, R. M., 1996. The end of the beginning: the race to begin human genome sequencing. *Genome Res.* **6** : 771-772.

Botstein, D., Chervitz, S. A., Cherry, J. M., 1987. Yeast as a model organism. *Science* **277** : 1259-1260.

Brumbaugh, J. A., Middendorf, L. R., Grone, D. L., Ruth, J. L., 1988. Continuous on-line DNA sequencing using oligodeoxynucleotide primers with multiple fluorophors. *Proc. Natl. Acad. USA* **85** : 5610-5614.

Buess, M., Moroni, C., Hirsch, H. H., 1997. Direct identification of differetially expressed genes by cycle sequencing and cycle labelling using the differential display PCR primers. *Nucleic Acids Res.* **25** : 2223-2235.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton,R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D. Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Venter, J. C., 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273** : 1058-1073.

Burke, D. T., Carle, G. F., Olson, M. V., 1987. Cloning of large segments of exogenous DNA into yeast by means of artifical chromosome vectors. *Science* **236** : 806-812.

Carothers, A. M., Urlaub, G., Mucha, J., Grunberger, D., Chasin, L. A., 1989. Point mutation analysis in a mammalian gene: rapid preparation of total RNA, PCR amplification of cDNA, and *Taq* sequencing by a novel method. *Biotechniques* **7** : 494-496

Chen, E. Y., Seeburg, P. H., 1985. Supercoil sequencing: a fast and simple method for sequencing plasmid DNA. *DNA* **4** : 165-170.

Chen, E. Y., Schlessinger, D., Kere, J., 1993. Ordered shotgun sequencing, a strategy for integrated mapping and sequencing of YAC clones. *Genomics* **17** : 651-656.

Chien, A., Edgar, D. B., Trela, J. M., *1976.* Deoxyribonucleic acid polymerase from extreme thermophile *Thermomus aquaticus*, *J. Bacteriol.* **127** : 1550-1557.

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., Rutter, W. J., 1979. Deoxyribonucleic acid from sources enriched in ribonuclease, *Biochemistry* **18** : 5294-9299.

Chomeczynski, P., Sacchi, N., 1987. Single-step method of RNA isolation by acid gunaidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162** : 156-159.

Chumakov, I. M., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billault, A., Guillou, S., Soularue, P., Guasconi, G., Poullier, E., Gros, I., *et al.*, 1995. A YAC contig map of the human genome. *Nature* **377** : 175-297.

Collins, F. S., 1995. Positional cloning moves from perditional to traditional. *Nat. Genet.* **9** : 347-350.

Collins, J., Hojn, B., 1978. Cosmids: a type of plasmid gene-cloning vector that is packageable *in vitro* in bacterophage lambda heads. *Proc. Natl. Acad. USA.* **75** : 4242-4246.

Coulson, A., Waterston, R., Kiff, J., Sulston, J., Kohara, Y., 1988. Genome linking with yeast artifical chromosomes, *Nature* **335** : *184-186.*

Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J., Waterston, R., 1991. YACs and C. elegans genome. *Bioassays* **13** : 413-417.

DeRisi, J. L., Iyer, V. R., Brown, P.O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278** : 680-686.

Drmanac, S., Stavropoulos, N. A., Labat, I., Vonau, J., Hauser, B., Soares, M. B., Drmanac, R., 1996. Gene-representing cDNA clusters defined by hybridization of 57419 clones from infant brain libraries with short oligonucleotide probe. *Genomics* **37** : 29-40.

Dudley, J. P., Butel, J. S., Socher, S. H., Rosen, J. M., 1978. Detection of mouse mammary tumor virus RNA in BALB/c tumor cell lines of nonviral etilogies. *J. Virol.* **28** : 743-752.

Efstratiadis, A., Kafatos, F. C., Maxam, A. M., Maniatis, T., 1976. Enzymatic in vitro synthesis of globin genes. *Cell* **7** : 279-288.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., 1995. Whole genome random sequencing and assembly of *heamophilus influenzae* Rd. *Science* **269** : 496-512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., 1995. The minimal gene complement of *Mycoplsma genitalium*, *Science* **270** : 397-403.

Forhman, M. A., Dush, M. K., Martin, G. R., 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA* **85** : 8998-9002.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., Claverie, J. M., 1998. Alternate polyadenylation in human mRNAs: a large-scale amalysis by EST clustering. *Genome Res.* **8** : 524-530,

Gingeras, T. R., Sciaky, D., Gelinas, R. E., Bing-Dong, J., Yen, C. E., Kelly, M. M., Bullock, P. A., Parsons, B. L., O'Neill, K. E., Roberts, R. J., 1982. Nucleotide sequences from the adenovirus-2 genome. *J. Biol. Chem.* **257** : 13475-13491.

Goffeau, A., et al., 1997. The yeast genome directory. *Nature* **387** (6632 Suppl.).

Green, P., 1997. Against a whole-genome shotgun. Genome Res. **7** : 410-417.

Gyllenstein, U. B., Josefsson, A., Schemschat, K., Saldeen, T., Petterson, U., 1992. DNA typing of foresic material with mixed genotype using allele-specific enzymatic amplification (polymease chain reaction). *Foresic Sci. Int.* **52** : 149-160.

Hara, E., Kato, T., Nakada, S., Sekiya, S., Oda, K., 1991. Substractive cDNA cloning oligo(dT)30-latex and PCR: isoation of cDNA clones specific to undifferentiated human embryoal

carcinoma cells. *Nucleic Acids Res.* **19** : 7079-7104.

Hauge, B. M., Goodman, H. M., 1992. In: Bckmann, J. S., Osborn, T. C. (Eds), Palnt genome: Methods for genetic and physical mapping. Kluwer, Dordrecht, The Nethrlands, pp. 101-139.

Henikoff, S., 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28** : 351-359.

Higuchi, R. G., Ochman, H., 1989. Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. *Nucleic Acids Res.* **17** : 5865.

Hibert, H., Himmelreich, R., Plagens, H., Herrmann, R., 1996. Sequence analysis of 56 kb from the genome of the bacterium *Mycoplasma pneumoniae* comparising the dnaA region, the atp operon and a cluster of ribosomal protein genes. *Nucleic Acids Res.* **24** : 628-639.

Hofte, H., Desprez, T., Amselem, J., Chiapello, H., Rouze, P., Caboche, M., Moisan, A., Jourjon, M. F., Charpenteau, J. L., Berthomieu, P., et al., 1993. An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. *Plant J.* **4** : 1051-1061.

Hoog, C., 1991. Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy. *Nucleic Acids Res.* **19** : 6123-6127.

Hornes, E., Korsnes, L., 1990. Magentic DNA hybridization properties of oligonucleotide probes attached to superparamagnetic beads and their use in the isolation of poly(A) mRNA from eukaryotic cells. *Genet. Anal. Tech. Appl.* **7** : 145-150.

Hubank, M., Schatz, D. G., 1994. Identifying differenves in mRNA expression by representational difference analysis of cDNA. *Nature Acids Res.* **22** : 5640-5648.

Hultman, T., Stahl, S., Horness, E., Uhlen, M., 1989. Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support. *Nucleic Acids Res.* **17** : 4937-4946.

Hultman, T., Bergh, S., Moks, T., Uhlen, M., 1991. Bidirectional solid-phase sequencing of in vitro-amplified plasmid DNA. *Biotechniques* **10** : 84-93.

Innis, M. A., Myambo, K. B., Gelfand, D. H., Brow, M. A., 1988. DNA sequencing with Thermus aaquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proc. Natl. Acad. Sci. USA* **85** : 9436-9440.

Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A., De Jong, P. J., 1994. A new bacterophage P1-ferived vector for the propagation of large human DNA fragments. *Nat. Genet.* **6** : 84-89.

Jakobsen, K, S., Breivold, E., Hornes, E., 1990. Purification of mRNA directly from crude palnt tissues in 15 minutes using magnetic oligo dT microspheres. *Nucleic Acids Res.* **18** : 3669.

Johnston, M., 1996. The complete code for a eukaryotic cell. Genome sequencing. *Curr. Biol.* **6** : 500-503.

Khurshid, F., Beck, S., 1993. Error analysis in manual and automated DNA sequencing. *Anal. Biochem.* **208** : 138-143.

Klenow, H., Henningsen, I., 1970. Selective elimination of the exonuclease activity of the deoxynucleic acid polymerase from E. coli B by limited protelysis. *Proc. Natl. Acad. USA* **65** : 168-175.

Ko, M. S., 1990. An 'equalized cDNA library' by reassociation of short double-stranded cDNAs. *Nucleic Acids Res.* **18** : 5705-5711.

Landsgren, U., Nilsson, M., Kwok, P. Y., 1998. Reading bits of genetic information: methods for single-nucleotides polymorphism analysis. *Genome Res.* **8** : 769-776.

Lee, L. G., Spurgeon, S. L., Heiner, C. R., Benson, S. C., Rosenblum, B. B., Menchen, S. M., Graham, R. J., Constantinescu, A., Upadhya, K. G., Cassel, J. M., 1997. New energy transfer dyrs for DNA sequencing. *Nucleic Acids Res.* **25** : 2816-2822.

Lennon, G., Auffray, C., Polymeropoulos, M., Soares, M. B., 1996. Consortium: an integrated molecular analysis of genome and their expression. *Genomics* **33** : 151-152.

Liang, P., Prdee, A. B., 1992. Differential display of eukaryotic messenger ENA by means of the polymerase chain reaction. *Science* **257** : 967-971.

Liew, C. C., Hwang, D. M., Fung, Y. W., Laurenssen, C., Cukerman, E., Tsui, S., Lee, C. Y., 1994. A catalog of genes in the cardiovascular system as identified by expressed sequence tags. *Proc. Natl. Acad. Sci. USA.* **91** : 10645-10649.

Lim, L., Canellakis, E. S., 1970. Adenine-rich polymer associated with rabbit reticulocyte RNA. *Nature* **227**: 710-712.

Loh, E. Y., Elliott, J. F., Cwirla, S., Lanier, L. L., Davis, M. M., 1989. Polymerase chain reaction with single specificity: analysis of T cell receptor delta chain. *Science* **243** : 217-220.

Marra, M. A., Hillier, L., Waterston, R. H., 1998. Expressed sequence tags-EST ablishing bridges between genomes. *Trend in Genet.* **14** : 4-7.

Maruyama, K., Sugano, S., 1994. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligonucleotides. *Gene* **138** : 171-174.

Maxam, A. M., Gilbert, W., 1977. A new method for sequencing DNA. *Proc. Natl. Acad. USA* **74** : 560-564.

McCombie, W. R., Adams, M. D., Kelley, J. M., FitzGerald, M. G., Utterback, T. R., Khan, M., Dubnick, M., Kerlavage, A. R., Venter, J. C., Fields, C., 1992. *Caenorhabditis elegans* expressed sequence tags identify gene families and potential diesase gene homologuous. *Nat. Genet.* **1** : 124-131.

Messing, J., Gronenborn, B., Muller-Hill, B., Hofschneider, P. H., 1978. Single-stranded filamentous DNA phage as a carrier for in vitro recombined DNA. In: Hofschneider, P. H., Starlinger, P. (Eds), Integration and Excision of DNA molecules, vol. 26. Springer, Berlin, pp: 29-32.

Misra, T. K., 1985. A new strategy to create orfered deletions for rapid nucleotide sequencing. Gene **34** : 263-268.

Moreno-Palanques, R. F., Fuldner, R. A., 1994. Automated DNA sequencing: Construction of cDNA libraries. Academic Press, London, pp: 102-1-109.

Murray, V., 1989. Improved double-stranded DNA sequencing using th linear polymerase chain reaction. *Nucleic Acids Res.* **17** : 8889.

Oefner, P., J., Hunicke-Smith, S. P., Chiang, L., Dietrich, F., Mulligan, J., Davis, R. W., 1996. Efficient random eubcloning of DNA sheared in a recirculating point-sink flow system. *Nucleic Acids Res.* **24** : 3879-3886.

Okubo, K., Yoshii, J., Yokouchi, H., Kameyama, M., Matsubara,

K., 1994. An expressed profile of active genes in human colonic mucosa. *DNA Res.* **1** : 37-45.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., Matsbara, K., 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* **2** : 173-179.

Olivar, S. 1996. A betwork approach to the systematic analysis of yeast gene function. *Trend in Genet.* **12** : 241-242.

Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., Frank, T., 1986. Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acads. USA* **83** : 7826-7830.

Patanjali, S. R., Parimoo, S., Weissman, S. M., 1991. Construction of a uniform-abundance cDNA library. *Proc. Natl. Acad. USA* **88** : 1943-1947.

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A., Baumeister, K., 1987. A system for rapid DNA sequencing with fluroscent chain-terminating dideoxynucleotides. *Science* **238** : 336-341.

Riles, M. A., Dutchik, J. E., Baktha, A., McCauley, B. K., Thayer, E. C., Leekie, M. P., Braden, V. V., Depke, J. E., Olson, M. V., 1993. Physical maps of the six smallest chromosomes of Sacchromyces cecevisiae at a resolution of 2.6 kilobase pairs. *Genetics* **134** : 81-150.

Ronaghi, M., Uhlen, M., Nyren, P., 1998. A sequencing method based in real-time pyrophosphate. *Science* **281** : 363.

Rounsley, S. D., Glodek, A., Sutton, G., Adams, M. D., Somerville, C. R., Venter, J. C., Kerlavage, A. R., 1996. The construction of Arabidopsis expressed tag assemblies. A new resource to facilitate gene identification. *Palnt Physiol.* **112** : 1177-1183.

Saiki, R. K., Scharf, S., Fallona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., Arnheim, N., 1985. Enzymativc amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230** : 1350-1354.

Sanger, F., Nicklen, S., Coulson, A. R., 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. USA* **74** : 5463-5467.

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., Peterson, G. B., 1982. Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* **162** : 729-773.

Sarker, G., Sommer, S. S., 1988. RNA amplification with transcript sequencing (RAWTS). *Nucleic Acids Res.* **16** : 5197.

Selleri, L., Eubanks, J. H., Giovannini, M., Hermanson, G. G., Romo, A., Djabali, M., Maurer, S., McElligott, D. L., Smith, M. W., Evans, G. A., 1992. Detection and cahracterization of chimeric yeast artifical chromosome clones by flurorecent in situ suppression hybridization. *Genomics* **14** : 536-541.

Shatkin, A. J., 1976. Capping of eukaryotic mRNA. *Cell* **9** : 645-653.

Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Boucherie, H., Mann, M., 1996. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. USA* **93** : 14440-14445.

Short, J. M., Fernandez, J. M., Sorge, J. A., Huse, W. D., 1988. Lambda ZAP: a bacteriophage lambda expression vector with in vivo excision propreties. *Nucleic Acids Res.* **16** : 7583-7600.

Smith, V., Brown, C. M., Bankier, A. T.,, Barrell, B. G., 1990. Seminautomated preparing of DNA templates for large-scale sequencing projects. *DNA seq.* **1** : 73-78.

Soraes, M. B., Bonald, M. F., Jelene, P., Su, L., Lawton, L., Efstratiadis, A., 1994. Construction and characterization of a nomalized cDNA library. *Proc. Natl. Acad. USA* **91** : 9228-9232.

Stahl, S., Hultman, T, Olsson, A., Morks, T., Uhlen, M., 1988. Solid phase DNA sequencing using the biotin-avidin system. *Nucleic Acids Res.* **16** : 3025-3038.

Sternberg, N., 1990. Bacteriphage P1 cloning system for the isolation, amplificatiom, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc. Natl. Acad USA* **87** : 103-107.

Strauss, F. W., Kobori, J. A., Siu, G., Sommer, S. S., 1986. Specific-primer-directed DNA sequencing. *Anal. Biochem.* **154** : 353-360.

Studier, F. W., 1989. A strategy for high-volumn sequencing of cosmid DNAs: ramdom and directed priming with a library of oligonucleotides. *Proc. Natl. Acad USA* **86** : 6917-6921.

Sulston, J. Du., Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., 1992. The C. elegans genome sequencing project: a begining. *Nature* **356** : 37-41.

Szybalski, W., 1990. Proposal for sequencing DNA using ligation of hexamers to generate sequencial elongation primers (SPL-6). *Gene* **90** : 177-178.

Tabor, S., Richardson, C. C. 1995. A single residue in DNA polymerases of the E. coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotide. *Proc. Natl. Acad. Sci. USA* **92** : 6339-6343.

Thierry, A., Gaillon, L., Galibert, F., Dujon, B., 1995. Construction of a complete genomic library of Saccharomyces cerevisiae and physical mapping of chromosome XI at 3.7 kb resolution. *Yeast* **11** : 121-135.

Tong, X., Smith, L. M., 1992. Solid-phase method for the purification of DNA sequencing reaction. *Anal. Chem.* **64** : 2672-2677.

Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., 1995. Serial analysis of gene expression. *Science* **270** : 484-487.

Venter, J. C. Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O., Hunkapiller, M., 1998. Shotgun sequencing of the human genome. *Science* **280** : 1540-1542.

Vieira, J., & Messing J., 1982. The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19** : 259-268.

Voss, H., Schwager, J., 1995. Efficient low redundancy large-scale DNA sequencing at EMBL. *J. Biotechnol.* **41** : 121-129.

Waterson, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green,P., Shownkeen, R., Halloran, N., 1992. A survey of expressed genes in Caenorhabditis elegans. *Nat. Genet.* **1** : 114-123.

Weber, J. L., & Myers, E. W., 1997. Human shole-genome shotgun sequencing. *Genome Res.* **7** : 401-409.

Welsh, J., et al., Arbitrarily primed PCR fingerfinting of RNA.

Nucleic Acids Res. **20** : 4965-4970.

Wiemann, S., Voss, H., Schwager, C., Rupp. T., Stegeman, J., Zimmermann, J., Grothues, D., Sensen, C., Erfle, H., Hewitt, N., 1993. Sequencing and analysis of 51.6 kilobases on th left arm of chromosome XI from Saccharomyces cerevisiae reveals 23 open reading frames including the FAS1 gene. *Yeast* **9** : 1343-1348.

Wilcox, A. S., Khan, A. S., Hopkins, J. A., Sikela., J. M., 1991. Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implification for an expression map of the genome. *Nucleic Acids Res.* **19** : 1837-1843.

Wilson, R. K., Koop, B. F., Chen, C., Halloran, J. A., Sikela, J. M., 1992. Nucleotide sequence analysis 0f 95 kb near the 3' end of the murine T-cell receptor alpha/delta chain locus: strategy and methodology. *Genomics* **13**: 1198-1208.

Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfiled, J., Connell, M., Copsey, T., Cooper, J., 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. *Nature* **368** : 32-38.

Zimmermann, J., Voss, H., Schwager, C., Stegemann, J., Erfle, H., Stucky, K., Kristenson, T., Ansorge, W., 1990. A simplified protocol for fast plasmid DNA sequencing. *Nucleic Acids Res.* **18** : 1067.