게놈 시대 인간 유전자 데이터베이스의 검색 및 사용법

강릉대학교 치과대학 및 구강과학연구소
최 영 님

ABSTRACT

Human Genome Database Mining in Post-Genome Era

College of Dentistry and Research Institute of Oral Science, Kangnung National University
Youngnim Choi, D.D.S., Ph.D.

The reports of human genome drafts last February opened a post-genome era. Now, a mine of information is waiting to be explored by the frontiersmen. In this article, various public databases tools and their application to biomedical research were introduced with an example analysis using oral squamous cell carcinoma.

인간 게놈 프로젝트

21C 첫 해 32 억 개의 염기 (3.2 Gigabases (Gb))로 구성된 인간의 총 유전정보 초안의 발표라는 역사적 사건과 함께 포스트-게놈 시대가 개막되었다. 인간 게놈 프로젝트는 24개 염색체 각각에 대해 하나로 연결된 단일 염기서열 (nucleotide sequence)을 밝히고 모든 유전자의 염색체 상 위치를 파악한다는 원대한 목표 하에 1990년 International Human Genome Sequencing Consortium이 구성되면서 시작되었다. 미국, 영국, 독일, 프랑스, 일본, 그리고 중국에 있는 20 개 실험실의 공동 노력은 마침내 지난 2월 15일 인간 게놈의 초안을 Nature에 발표하기에 이르렀고[1], 그 모든 정보는 인터넷을 통해 무료로 접근할 수 있다[2]. 아직 인간 게놈이 완성되지 않은 상황에서는, 비록 사용료를 지불해야 하지만, 동시에 Science에 발표된 Celera Genomics사에서 수행한 게놈의 초안도 유용할 것이다[3].

그럼, 지난 2월에 발표된 게놈 초안에서 어떤 것을 얻을 수 있는가 살펴보자. 아직 완성되지 않은 초

본 논문은 2001년도 복지부 보건의료기술연구개발사업 01-PJ5-PG3-20507-0010로 지원되었음.

안은 인간 게놈의 약 90%에 달하는 염기서열을 해독한 것인데, 아직 하나로 연결된 단일 염기서열이 아니라 1,000 개 이상의 단절 부위 (discontinuities)를 포함하고 있다. 그러나, 이 정도의 정보로도 많은 의학 연구에 유용하게 이용될 수 있고 인간의 게놈이 어떻게 구성되어 있는지에 대해 중요한 단서를 제공하고 있다[1]. 32억 개의 염기 중 실제로 단백질 합성을 지시하는 정보를 담고있는 부위는 1.1~1.4%에 불과한데, 이것은 RNA로 전사되는 부분의 5~28%에 해당된다. 포유류 염색체의 놀라운 특징 중의 하나는 무척추 생물인 연충이나 초파리와 달리 염색체상의 유전자 분포가 균일하지 않다는 점이다. 우리의 염색체에는 많은 유전자가 고밀도로 모여 있는 부위가 있는가 하면, 소위 "junk" DNA만이 발견되는 사막과 같은 부위도 있다. Junk DNA는 여러 가지 종류의 반복염기서열 (repeating sequences)로 구성되어 있는데, 이러한 junk DNA가 우리 게놈의 무려 50%를 차지하고 있다. 이것은 지금까지 게놈이 완전히 밝혀진 mustard weed(11%), 연충(7%), 초파리(3%)에 비해 매우 높은 비율로, 이러한 반복염기서열이 그야말로 junk인지 아니면 새로운 중요한 기능을 하는지는 앞으로 연구해야 할 과제이다. 두 번째로 놀라운 사실은 사람의 총 유전자 수가 초파리의 불과 두 배에 불과한 32,000~39,000개로 추산된다는 점이다. 비록 총 유전자 수는 예상보다 많지 않지만 하나의 사람 유전자는 평균 3개의 다른 단백질을 생성할 수 있고, 따라서 사람의 게놈이 생성하는 총 단백질 집합(proteom)은 무척추 생물의 경우에 비해 훨씬 복잡하다.

International Human Genome Sequencing Consortium이 발표한 게놈 초안은 염기서열 뿐 아니라 BACs 지도와 SNPs (Single nucleotide polymorphism) 지도라는 중요한 별책부록을 동반하고 있다.

많은 동일한 반복 염기서열이 존재하는 염색체의 단일 염기서열을 얻기 위해, 인간게놈프로젝트는 클론에 기초한 방법을 사용하였다. 즉, 32억 개의 염기로 구성되는 게놈을 제한효소를 이용해 약 150,000개의 염기로 구성되는 DNA 조각들로 자른 후, 각 DNA 조각은 Bacterial Artifitial Chromosome (BAC)에 집어넣어 세균 (bacteria)으로 하여금 증폭하도록 함으로써 동일한 DNA 조각을 만들어 내는 cloning 과정을 거치게 된다.

이렇게 만들어진 각각의 BAC 클론은 더 작은 많은 DNA 조각으로 잘라 다시 cloning 한 후, 작은 DNA 조각들의 염기서열을 분석해 조합함으로써 BAC clone 전체의 염기서열이 만들어지게 된다. 그리고 BAC 클론들을 이어 붙임으로써 염색체의 단일 염기서열을 얻는다. Consortium은 염기서열을 분석하기 전에, fluorescence in situ hybridization (FISH) 방법을 이용해 30,000 개 BAC clones이 각 염색체의 어느 부분(밴드로 표시)에 위치하는지 먼저 파악하고, 지도 (Cytogenetic map)를 작성하였다[4]. 많은 질병에 대해 병인에 관여하는 연관유전자부위 (linked gene loci)가 연관분석 (linkage analysis)을 통해 이미 연구되어 있기 때문에, 이제는 BAC 유전자 지도에서 연관부위에 위치하는 BAC 클론을 찾고, 그 클론의 염기서열로부터 질병에 관여하는 유전자를 찾는 작업이 훨씬 가속화될 것이다. SNPs 지도는 질병과 연관된 유전자의 역할을 연구하는데 있어서 또 하나의 귀중한 자료이다[5]. SNP란 개체간의 게놈 염기서열을 비교해 하나의 염기가 서로 상이함을 보이는 부위를 말한다. SNP는 개체간 유전자 변이의 가장 흔한 형태인데, 유전자 변이는 한 개인이 특정 질병에 걸릴 확률, 걸린다면 언제 발병하여 얼마나 심한 형태로 앓을지, 그리고 특정 치료약에 어떻게 반응할지 등 질병과 관련된 여러 양상 뿐 아니라, 개체간에 관찰되는 다양한 신체적, 생리적 특성에 대한 유전적 배경을 반영한다.

SNP 지도는 전 염색체에 걸쳐 평균 1.9kb마다 하나씩 관찰되는 SNP 142만 개를 찾아내 염색체 상의 위치를 명기한 지도이다. 142만 개의 SNPs 중 60,000 개는 유전자를 coding하는 exon 상에 위치하며, 95%의 유전자가 적어도 하나의 SNP를 가지고 있다. 환자와 대조군에서 전 게놈에 걸쳐 분포하는 SNP 빈도를 비교함으로써, 어떤 SNP가 어떤 질병과 연관되는지 알 수 있고, 이것은 다시 어떤 유전

자가 질병에 관여하는지 알 수 있다. 이와 같은 SNP 지도가 의학연구에 유용하려면 데이터베이스에 있는 SNPs가 염기서열분석 과정의 에러가 아닌 진정한 다형성이어야 하며 SNPs를 찾기 위해 사용된 시료 뿐 아니라 다른 시료에서도 다형성을 보여야한다. SNP working group이 142만 개 중 1,585개의 SNPs를 골라 좀 더 확인해 본 결과 95%가 진정한 SNPs 였으며, 1,276 개의 SNPs를 더 많은 시료에 대해 조사해 본 결과 적어도 82%가 다형성을 보였다.

이것은 발표된 SNPs의 유용성을 제시하는 것으로, 나머지 SNPs에 대해서도 앞으로 연구를 통해 기초자료를 확충해 나갈 필요가 있다. 특히, SNP 중에서도 두 개 대립형질이 모두 높은 빈도로 나타나는 것이 질병 연구에 가장 유용하기 때문에 각 SNP의 특정 인구집단에서의 빈도에 대한 자료를 확충해나가는 것이 매우 중요하다.

따라서 현 시점에서 한국에서 조속히 해야 할 과제는 한두 명 한국인의 게놈 염기서열 (어차피 발표된 것과 99.9% 동일할 것이다)을 알아내는 게 아니라 한국인에서의 SNPs 빈도를 조사하는 일이다.

포스트-게놈 시대의 연구 방향

게놈프로젝트와 함께 최근 생의학 연구에 혁명적 변화를 가져온 것은 게노믹스 (genomics)와 프로테오믹스 (proteomics) 같은 새로운 기술이다. 한 번에 관찰할 수 있는 유전자 수가 종래의 수 개에서 수천 수만 개로 늘어남으로써, 특정 세포나 조직의 총체적 유전자 발현 양상을 체계적으로 관찰할 수 있게 된 것이다. 최근 cDNA microarray나 high-density oligonucleotide array와 같이 게노믹스 테크닉을 이용해 유전자 발현 양상을 연구한 논문들이 보여주듯이 이제는 질병이나 다양한 환경요인 등 특정 요소에 의해 조절되는 모든 유전자의 목록을 획득할 수 있다.

그러나, 유전자 발현에 대한 정보는 연구의 시작일 뿐이고, 궁극적으로는 다시 하나의 유전자로 돌아가 발현된 유전자가 질병의 병인에 관여하는지, 관여한다면 생물학적 과정 (biological process)에서 구체적으로 어떤 기능을 하는지 밝혀야 한다. SNPs 지도에 있는 각 유전자의 SNPs 정보를 사용해 질병과 유전자 변이간의 연관을 연구할 수 있으며, 단백질의 기능은 다양한 창의적인 생물학적 실험을 통해 밝힐 수 있다.

게노믹스와 프로테오믹스가 요즘 유행하는 첨단 연구기법이나, 필요한 장비와 비용이 매우 비싸기 때문에 일반적인 실험실에서 하기에는 적절하지 않다. 이에 비해 유전자 발현을 연구하는 전통적인 방법들은 시간과 노력이 많이 소요된다.

본 저자는 데이터베이스에 있는 유전자 발현 데이터와 새로운 게놈 데이터를 한번 이용해 볼 것을 권한다. 모든 사람이 같은 선에서 출발하기보다는 이미 되어있는 유전자 발현 연구결과에서 흥미 있는 유전자를 골라 다음 단계의 연구에 들어갈 수 있다.

공공 도메인에 있는 유전자 발현 데이터 (Gene Expression Data in public domain)

유전자 발현 양상에 대해 가장 체계적이고 방대한 데이터를 축적하고 있는 곳은 미국 National Center for Biotechnology Information (NCBI)에서 제공하는 Cancer Genome Anotomy Project (CGAP) 데이터베이스이다.

CGAP은 미국 국립 암연구소 (National Cancer Institute)가 암세포의 분자생물학적 특성을 해부하는데 필요한 정보를 축적하고 정보 분석에 필요한 도구들을 개발하기 위해 1996년 시작한 사업으로, 그 목표는 정상, 전구암, 그리고 암세포의 유전자 발현 양상을 분석하여 궁극적으로 암의 조기진단과 치료를 향상시키고자 하는 것이다[6].

CGAP에 있는 유전자 발현 데이터는 EST 데이터와 SAGE 데이터 두 가지 종류가 있다. EST란 Expressed Sequence Tags의 축약어로 cDNA library에서 무작위로 선택한 클론의 single path sequencing (염기서열분석을 한번만 시행하는 것)으

로 얻은 유전정보이다.

따라서 보통 200~500 bp에 달하는 EST는 단백질 코딩에 대한 정보는 없는 경우도 많지만 전사되는 각 유전자 고유의 서명과 같기 때문에, 조직에서 추출한 mRNA에서 cDNA library를 만든 후 자동염기서열분석기를 이용해 많은 클론들을 대량으로 분석해 EST 데이터를 얻음으로써 비교적 간단하게 그 조직에서 발현되는 유전자 종류와 양을 가늠할 수 있다. CGAP에서 생성되는 모든 EST 데이터는 EST 데이터베이스인 dbEST에 통합되며, 모든 EST clone은 최소한의 비용으로 American Type Culture Collection, Incyte Genomics, Research Genetics Incorporation, UK Human Genome Mapping Project Resource Center, 또는 Resource Center of the German Human Genome Project에서 구입할 수 있다. 또한 CGAP cDNA library도 Stratagene이나 Life Tecnhologies 사에서 구입하거나 NIH의 Dr. Robert Straousberg를 통해 얻을 수 있다[7]. 그러므로, EST 데이터의 장점은 유전자 발현 비교 결과 의미 있는 유전자를 발견했을 때, 바로 다음 단계 연구에 착수할 수 있도록 클론된 유전자 조각을 얻을 수 있다는 점이다. NCBI는 매주 새로운 library의 데이터를 첨가하고 있는데, 2001년 8월 7일 현재 CGAP 싸이트(cgap.nci.nih.gov)에 축적되어 있는 사람의 EST library는 총 6873 개로, 이들은 51가지 조직유형으로 분류되어 있다. 이 중 886개가 두경부 (Head and Neck) 영역에 발생한 암종(carcinoma)을 포함한 두경부 조직의 cDNA library이다.

SAGE란 Serial Analysis of Gene Expression의 축약어로 EST 대신 10 bp만을 유전자 식별을 위한 고유 tag로 사용함으로써, 자동염기서열분석을 이용해 세포에서 발현되는 유전자 양상을 정량화 하는 능력을 향상시킨 방법이다[8].

SAGE 데이터는 특정 세포에서 발현된 tags 종류와 각 tag의 수로 표현되는데, SAGE에서 사용되는 tag는 전사된 cDNA에서 특정 제한효소로 인지되는 곳 중 가장 3′에 위치하는 지점에서 3′ 쪽에 있는 9~10 bp로 구성된다. 짧은 tag를 사용함으로써 분석 속도를 높이고 분석비용을 줄인 반면, 각 tag가 유전자를 지정하는 정확성, 따라서 유전자 발현 정량화의 정확성은 다소 떨어진다.

그러나, 보통 수백 개의 클론을 분석하는 EST 데이터에 비해 십만 개 이상의 tags를 분석하는 SAGE 데이터는 유전자 발현 양상을 숫자화 할 수 있다는 장점을 갖고 있다. SAGE 데이터 역시 계속 늘어나고 있는데, 2001년 8월 7일 현재 SAGE 싸이트 (www.ncbi.nlm.nih.gov /SAGE)에는 14가지 유형의 조직에서 얻어진 169개 library의 데이터가 축적되어 있다.

NCBI는 1999년 유전자 발현 데이터의 보급과 공공 사용을 돕기 위해 Gene Expression Omnibus 사업을 발족하여, 2000년에 GEO 싸이트를 개설하였다. GEO 싸이트 (www.ncbi.nlm.nih.gov/geo)는 다양한 생물에서 연구된 유전자 발현 데이터의 저장고를 구축하고자 하는 노력으로, SAGE 뿐 아니라 spotted microarray, high-density oligonucleotide array, hybridization filter, 등 다양한 새로운 방법으로 얻어진 데이터를 모아놓고 있다. 2001년 8월 28일 현재 올라 있는 데이터는 570개 시료에 불과하나, 그 수는 빠른 속도로 증가할 것이므로 앞으로 많은 유용한 정보를 얻을 수 있을 것이다.

CCAP의 검색 및 비교 분석 프로그램

NCBI는 CGAP 데이터베이스를 검색하고 유전자 발현을 비교, 분석할 수 있는 다양한 도구 프로그램을 개발하여 제공하고 있다. 먼저 프로그램의 종류와 기능을 간단히 소개한 후, 실례를 들면서 사용법을 좀 더 자세히 소개하도록 하겠다. 이하 모든 도구는 cgap.nci.nih.gov/Tools에서 연결된다.

1. 유전자 검색 도구

Gene Finder : 다양한 검색 기준에 따라 유전자를 검색하여 NCBI와 NCI의 각종 데이터베이스에 있는 유전자 정보에 연결시켜준다.

GO Browser : 사람과 생쥐 유전자를 기능, 생물학적 과정, 그리고 세포 성분에 따라 분류한다.

Nucleotide BLAST : 염기서열의 유사성을 바탕으로 유전자를 검색한다.

2. cDNA Library 검색 도구

Library Finder : 정해진 조건에 따라 데이터베이스를 검색해 조직 특이적인 library를 찾아내고, 각 library에 대한 자세한 정보를 제공한다.

3. 유전자 발현비교를 위한 도구

Gene Library Sorter (GLS) : 하나 또는 그룹으로 지정한 library에서 발현되는 모든 유전자를 찾아, 지정한 library에서만 발현되는 고유 유전자 (unique gene)와 다른 library에서도 발현되는 비고유 유전자 (non-unique gene)으로 분류하며 이들 고유와 비고유 유전자는 다시 알려진 것과 알려지지 않은 것으로 분류하여 유전자 목록을 작성한다.

cDNA xProfiler : 두 그룹의 cDNA library간에 유전자 발현을 비교한다. 두 그룹에 공통으로 발현되는 유전자와 어느 한쪽에만 발현되는 유전자를 GLS에서처럼 고유, 비고유, 알려진 것, 그리고 알려지지 않은 것으로 분류하여 보여준다.

Digital Gene Expression Displayer (DGED) : 두 그룹 library의 유전자 발현 양상을 비교해 통계적으로 유의한 차이가 있는 유전자를 찾아낸다.

SAGEmap xProfiler : 선택된 SAGE library간의 유전자 발현 양상을 비교한다.

SAGEmap vNorthern : SAGE tags를 동정하고 유전자 발현 양상을 수량화한다.

4. 염색체 관련 도구

Mitelman Database : 사람의 종양에서 발생하는 염색체 breakpoint에 대한 증례보고들을 모아 데이터베이스를 구축하였다.

Recurrent Aberrations : 종양에 빈번하게 나타나는 염색체 변이를 Mitelman 데이터베이스에서 검색한다.

FISH-mapped BACs : BAC 클론의 염색체 상의 위치를 검색한다.

5. SNPs 검색 도구

Expression-Based SNP Imagemaps : 종양 유형과 염색체에 따라 SNPs를 검색할 수 있다.

Genetic and Physical SNP Maps : 각 SNP의 염색체 상 위치를 검색한다.

EST 데이터베이스와 분석 프로그램 이용의 실례 : 정상 구강 편평상피 (Squamous epithelium)와 편평세포암종 (Carcinoma)의 유전자 발현 비교

정상 구강 상피세포와 편평세포암종의 유전자 발현을 비교하여 구강 내 암종 발생에 어떤 유전자들이 관여하는지 찾기 위해 CGAP 데이터베이스에 있는 데이터를 검색 분석해 보았다.

1) Library Finder를 이용한 library 검색

Library Finder 검색 엔진은 검색 효율을 높이기 위해 몇 가지 변수를 선택하도록 되어 있다. Organism에서 Homo sapience와 Mus musculus 중 하나를 선택하고, Library Group으로 CGAP libraries, MGC libraries, ORESTES libraries, All EST libraries, SAGE libraries 중 하나를 선택할 수 있으며, Tissue Type으로 51가지 유형 중에 하나를 선택하거나 (Any)를 선택함으로써 모든 조직의 library를 모두 검색하도록 할 수도 있다. Tissue Preparation은 microdissected, bulk, cell line, flow sorted, 또는 (Any)를 선택할 수 있다. bulk는 통상적인 외과적인 수술로 떼어낸 조직이기 때문에 다양한 종류의 세포로 구성된 조직이 발현하는 유전자 library인데 반해, 현미경 하에서 레이저로 특정 세포만 미세하게 떼어낸 microdissected, FACS로 단일 종류의 세포만 분리해낸 flow sorted, 세포주를 사용한 cell line의 경우는 한 가지 세포가 발현하는 유전자 library이다. Tissue histology는 Normal, Precancer, Cancer 중 하나를 선택하거나 역시 (Any)

표1. 두경부 영역에서 미세해부기법으로 작성된 유전자 발현 library 목록.

Title	Tissue	Histology	Type	Protocol	현재까지 분석한 클론 수/총 클론 수
NCI_CGAP_HN7	Head and neck (Floor of mouth squamous epithelium)	normal	Microdissected	Non-normalized Krizman protocol1	(249/1536)
NCI_CGAP_HN8	Head and neck (Floor of mouth)	cancer (invasive carcinoma)	Microdissected	Non-normalized Krizman protocol1	(271/5376)
NCI_CGAP_HN9	Head and neck (Retromolar trigone squamous epithelium)	normal	Microdissected	Non-normalized Krizman protocol1	(1063/6144)
NCI_CGAP_HN10	Head and neck (Retromolar trigone)	cancer (carcinoma in situ)	Microdissected	Non-normalized Krizman protocol1	(582/42240)
NCI_CGAP_HN11	Head and neck (Tongue squamous epithelium)	normal	Microdissected	Non-normalized Krizman protocol1	(706/1536)
NCI_CGAP_HN12	Head and neck (Tongue)	cancer (invasive carcinoma)	Microdissected	Non-normalized Krizman protocol1	(725/4224)
NCI_CGAP_HN16	Head and neck (Retromolar trigone)	cancer (invasive carcinoma)	Microdissected	Non-normalized Krizman protocol1	(115/4608)

를 선택할 수도 있다. Library protocol은 크게 non-normalized, normalized, subtracted로 분류된다. Normalized란 library에서 클론을 무작위로 골라 염기서열분석을 할 때, 발현빈도가 높은 유전자만 반복적으로 선택되고 발현빈도가 낮은 유전자는 분석되지 않는 현상을 막기 위해 유전자들의 발현빈도를 표준화 한 후 library 제작을 한 것이다.

따라서 특정 조직에서 발현되는 모든 유전자를 알고 싶다면 normalized를 선택하는 게 좋고, 두 개의 조직에서 유전자 발현 빈도를 비교하고 싶다면 non-normalized를 선택하는 게 좋다. Subtracted란 미리 비교하고자 하는 대상의 mRNA를 사용해 공통되는 유전자들을 제거한 후 library 제작을 한 것이다. Library protocol도 (Any)를 선택할 수 있다. 만약 검색하고자 하는 library 이름을 정확히 안다면 Library Name 난에 써서 바로 검색할 수도 있다. Homo sapience, All EST libraries, head and neck, microdissected, (any)와 같은 변수를 넣고 검색한 결과 표 1과 같은 library 목록을 얻었다.

Title의 각 library 이름을 클릭하면 library의 보다 상세한 정보를 알려주는 페이지로 연결되는데, 도움이 되는 추가정보를 표 1에 괄호 안에 첨가하였다.

2) cDNA xProfiler를 이용한 유전자 발현 비교

cDNA xProfiler 역시 Library Finder 프로그램과 같은 변수를 선택하되 비교하고자 하는 Pool A와 Pool B를 각각 지정하게 되어있다. Pool A에는 Head and neck, microdissected, normal을 Pool B에는 Head and neck, microdissected, cancer의 변수를 넣고 검색하면 표 2에서와 같이 7개 library를 포함한 목록이 나온다. 목록의 library를 모두 선택하여 비교할 수도 있지만, 같은 조직에서 암종 발생에 의해 변화되는 유전자들을 살펴보기 위해 Pool A에는 NCI_CGAP_HN11만 선택하고 Pool B에는 NCI_CGAP_HN12만 선택한 후 다음 검색 과정을 실행하였다. NCI_CGAP_HN11과 NCI_ CGAP_HN12의 유전자 비교 결과는 다음과 같은 표로 정리된다 (표 3).

여기에서 A unique gene이란 A (B에서도 발견될 수 있음)에서 발견되지만 A와 B를 제외한 다른 어떤 library에서도 발견되지 않는 유전자를 말하며, A muniu B unique gene이란 오직 A에서만 발견되는

표 2. Pool A and B Setup for Expression XProfiler

Pool A	Pool B	Library Name	Sequences	Keywords
		NCI_CGAP_HN7	249	head and neck, epithelium, normal, mouth, EST, CGAP, non-normalized, Krizman protocol 1, microdissected
		NCI_CGAP_HN8	271	head and neck, invasive, carcinoma, EST, CGAP, non-normalized, Krizman protocol 1, microdissected
		NCI_CGAP_HN9	1063	head and neck, epithelium, normal, EST, CGAP, non-normalized, Krizman protocol 1, microdissected
		NCI_CGAP_HN10	582	head and neck, carcinoma, EST, CGAP, non-normalized, Krizman protocol 1, microdissected
V		NCI_CGAP_HN11	706	Head and neck, epithelium, normal, tongue, EST, CGAP, non-normalized, Krizman protocol 1, microdissected
	V	NCI_CGAP_HN12	725	Head and neck,tongue, invasive, poorly differentiated, moderately differentiated, carcinoma, EST, CGAP, non-normalized, Krizman protocol 1, microdissected
		NCI_CGAP_HN16	115	head and neck, carcinoma, EST, CGAP, non-normalized, Krizman protocol 1, microdissected

유전자를 의미한다.

A non-unique gene이란 A (B에서도 발견될 수 있음)에서 발견될 뿐 아니라 A와 B를 제외한 다른 library에서도 발견되는 유전자를 뜻하고, A muniu B non-unique gene이란 A와 다른 library에서 발견되지만 B에서는 발견되지 않는 유전자를 뜻한다. Known gene은 유전자의 염기서열이 완전히 알려진 것이고, unknown gene은 유전자 coding 여부가 알려지지 않은 것이다. 각 항의 숫자를 클릭하면 다음과 같은 유전자 목록을 볼 수 있다. 일례로 표 4에 혀의 편평상피와 혀에 발생한 편평상피암종에서 공통으로 발현된 16개의 non-unique known gene을 정리하였다.

각 유전자의 Gene Info를 클릭하면 EST 데이터로부터 유전자가 발현되는 조직 분포, 염색체 상 위치, 다른 유사 단백질에 대한 정보를 얻을 수 있으며, UniGene (gDNA, mRNA, EST, protein 등 발표된 모든 sequence 통합), LocusLink (염색체 상 위치와 linked marker에 관한 정보), OMIM (Online Mendelian Inheritance in Man : 사람의 유전자와 관련된 질병에 대한 고찰), SNPs (유전자 주변의 SNPs 검색) 등의 데이터베이스로 연결되어 유전자에 대한 다양한 정보를 얻을 수 있다. 정상 상피와 암종에서 공통으로 발현된 유전자는 ribosomal protein이 가장 많았고, kertin 6A와 desmoplakin을 포함해 IgE에 반응해 히스타민 분비 인자로 작용하는 TPT1처럼 상피세포 특성과 관련된 유전자, 그밖에 항원 표지와 관련된 beta-2-microglobulin, calnexin 등이 있었다.

혀의 정상 편평상피 (NCI_CGAP_HN11)와 다른 library에서 발견되었지만 편평상피암종 (NCI_CGAP_HN12)에서는 발견되지 않은 104 개의 알려진 유전자 목록에서 Catenin delta-1은 주목할 만했다. Catenin은 세포막 단백질인 cadherin의 cytoplasmic domain에 연결되는 단백질로 EGF receptor, PDGF receptor, 그리고 CSF receptor를 통한 신호전달 뿐 아니라 src에 의한 cell transformation에 관여한다. Cadherin/catenin 복합체의 다른 구성요소처럼 delta-catenin의 결함도 악성종양에 기여할 가능성이 있는데, Dillon 등은 침습성 유방관암종 (invasive ductal breast carcinoma)의 약 10%에서 delta-catenin의 발현이 완전히 상실된 것을 보고한 바 있다[9]. 따라서 혀의 편평상피암종의 발생에도 관여할 가능성이 높다.

표 3. XProfiler Results (Libraries in A: NCI_CGAP_HN11; Libraries in B: NCI_CGAP_HN12)

Subset	Unique Genes		Non-Unique Genes	
	Known	Unknown	Known	Unknown
A	0	32	120	56
B	0	24	205	52
A Or B	0	56	309	104
A and B	0	0	16	4
A minus B	0	32	104	52
B minus A	0	24	189	48

표 4. XProfiler Genes (A and B, Known, Non-unique)

Symbol	Name	Sequence ID	CGAP
ARL6IP	ADP-ribosylation factor-like 6 interacting protein	D31885	Gene Info
B2M	beta-2-microglobulin	NM_004048	Gene Info
CANX	calnexin	NM_001746	Gene Info
DSP	desmoplakin (DPI, DPII)	NM_004415	Gene Info
EIF2S3	eukaryotic translation initiation factor 2, subunit 3	NM_001415	Gene Info
HRIHFB2122	putative nuclear protein	NM_007032	Gene Info
KRT6A	keratin 6A	NM_005554	Gene Info
RAB7	RAB7, member RAS oncogene family	NM_004637	Gene Info
RPL12	ribosomal protein L12	NM_000976	Gene Info
RPL13A	ribosomal protein L13A	NM_012423	Gene Info
RPL39	ribosomal protein L39	NM_001000	Gene Info
RPL44	ribosomal protein L44	NM_021029	Gene Info
RPL5	ribosomal protein L5	NM_000969	Gene Info
RPS3	ribosomal protein S3	NM_001005	Gene Info
TPT1	tumor protein, translationally-controlled 1	NM_003295	Gene Info
UBA52	ubiquitin A-52 residue ribosomal protein fusion product 1	NM_003333	Gene Info

반대로 혀의 편평상피암종 (NCI_CGAP_HN12)과 다른 library에서 발견되었지만 정상 편평상피 (NCI_CGAP_HN11)에서는 발견되지 않은 189 개의 유전자 목록에는, 암종에 특이한 유전자는 아니라도 세포의 빠른 증식과 높은 대사를 반영하는 유전자들이 많았는데, 즉 H3 histone 3A와 3B, histone deacetylase 3, helicase-moi, topoisomerase II alpha와 같이 DNA 복제나 RNA 전사에 관여하는 유전자, 단백질 합성에 필요한 17 종의 ribosomal protein, 그리고 세포 증식에 필수적인 철이온을 운반하고 저장하는 transferrin receptor와 ferritin이 있었다. 그 밖의 유전자들 중에는 종양의 증식과 침습에 밀접한 관계가 있는 High Mobility Group Protein 1 (HMG 1)이 있었다.

이와 같은 유전자 발현 비교는 cDNA library의 모든 클론이 세포질의 각 mRNA 분자를 대표한다는 가정에 근거하지만, 실제로 분석된 700여 개의 클론은 하나의 세포가 발현하는 총 mRNA의 극히 일부에 지나지 않기 때문에, 관찰된 유전자 발현 차이가 유의한 것인지 아니면 단순히 충분한 클론을 분석하지 않았기 때문인지 의문이 남는다. 이러한 문제점을 보완하는 분석 도구가 Digital Gene Expression Displayer이다.

표 5. DGED Results
Total Sequences in Pool A : 244
Total sequences in Pool B : 322
Total libraries in Pool A : 1 (NCI_CGAP_HN11)
Total libraries in Pool B : 1 (NCI_CGAP_HN12)
p-value filter : 0.1

Symbol	Gene Name	aAccession	Libraries		Sequences		Seq Odds	Chi Squ
			A	B	A	B	A:B	P<
S100A8	S100 Calcium-Binding Protein A8	NM_002964	1	0	6	0	NaN	4.67e-03
KRT5	Keratin 5	NM_000424	1	0	4	0	NaN	2.11e-02
-	EST	BG573828	1	0	4	0	NaN	2.11e-02
-	EST	AW275730	1	0	3	0	NaN	4.60e-02
RPL37A	Ribosomal Protein L37A	NM_000998	0	1	0	4	0	8.06e-02
ACTB	Actin Beta	NM_001101	0	1	0	4	0	8.06e-02
GTF2I	General Transcription Factor 2I	NM_001518 NM_032999 NM_033000 NM_033001 NM_033003	0	1	0	4	0	8.06e-02

NaN : Not a number, 분모인 B가 0인 경우.

3) Digital Gene Expression Displayer (DGED)를 이용한 유전자 발현 비교

단순히 유전자 유무를 비교하는 cDNA xProfiler에 비해, DGED는 특정 유전자가 지정한 library pool A에서 관찰되는 정도와 library pool B에서 관찰되는 정도를 비교하여 유전자 발현 차이의 유의성을 Chi-square 테스트로 검증한다.

Chi-square 테스트는 pool A에서 발견되는 특정 유전자 a의 수와 a를 제외한 모든 유전자 수의 비(다시 말해 발현된 전체 유전자 집합에서 a가 차지하는 비율)가 pool B에서 발견되는 특정 유전자 a의 수와 a를 제외한 모든 유전자 수의 비와 같은지를 물어 그 p 값이 0.05 이하일 때 pool A와 pool B에서 유전자 a의 발현 정도가 다르다고 할 수 있다.

사용방법은 cDNA xProfiler와 같은 요령으로 변수를 지정해 먼저 library 목록을 얻는다. Pool A와 pool B의 library를 선택한 후 두 가지 변수, 즉, pool A에 있는 유전자의 최소 숫자와 p 값을 설정하고 검색을 의뢰하면 된다.

첫 번째 변수인 Pool A에 있는 유전자의 최소 숫자는 기본 값이 2로 되어 있는데, 그 의미는 검색 결과에 pool A에서 최소한 2개 이상 발견된 유전자들만 포함한다는 뜻이다. 만약 pool A에는 발현되지 않고 pool B에서 많이 발현되는 유전자에 대한 정보도 모두 얻고 싶다면 이 값을 0으로 설정하는 게 좋다. p값의 기본 값은 0.01로 되어 있고, 그 의미는 검색 결과 p 값이 0.01보다 큰 것은 제외한다는 뜻이다. 보통 p 값이 0.05보다 작을 때 유의한 것으로 해석하므로 0.05로 설정하는 게 좋고, 유의성이 떨어지더라도 더 많은 유전자를 얻고 싶다면 더 큰 값으로 설정할 수도 있다.

0과 0.1의 변수를 사용하여 NCI_CGAP_HN11과 NCI_CGAP_HN12의 유전자 발현 비교를 한 결과 표 5와 같은 결과를 얻었다. 여기에서 total sequence 수는 표 2에서처럼 분석된 클론의 총 수가 아니라 분석 결과 얻어진 유전자의 총 가짓수이다.

표 5에서 library와 sequence 항목의 숫자는 각 pool에서 특정 유전자가 발견된 library 수와 sequence 수를 각각 나타낸다. S100A8, KRT5, 그리고 두 개의 EST 등 모두 4 개의 유전자가 정상 상피에는 발현되나 편평상피암종에서 발현되지 않았으며 p 값은 모두 0.05보다 작았다. 유의성 ($p<0.0806$)

표 6. DGED Results
Total Sequences in Pool A : 815
Total sequences in Pool B : 455
Total libraries in Pool A : 3
Total libraries in Pool B : 3
p-value filter : 0.05

Symbol	Gene Name	aAccession	Libraries		Sequences		Seq Odds	Chi Squ
			A	B	A	B	A:B	P<
-	EST	AW238182	1	0	181	0	NaN	1.88e-27
KRT4	Keratin 4	X07695	2	0	19	0	NaN	1.03e-03
RPS27A	Ribosomal Protein 27A	NM_002954	1	1	15	1	8.51	1.30e-02
-	EST	AW238163	1	0	9	0	NaN	2.45e-02
KRT13	Keratin 13	NM_002274	2	0	8	0	NaN	3.40e-02
-	EST	AW238421	1	0	7	0	NaN	4.74e-02
GTT1		NM_020151	1	0	7	0	NaN	4.74e-02
-	EST	AW438529	1	0	7	0	NaN	4.74e-02
S100A8	S100 Calcium-Binding Protein A8	NM_002964	3	2	41	11	2.14	2.42e-02
RPL37A	Ribosomal Protein L37A	NM_000998	0	1	0	4	0	7.34e-03
ACTB	Actin Beta	NM_001101	0	1	0	4	0	7.34e-03
GTF2I	General Transcription Factor 2I	NM_001518 NM_032999 NM_033000 NM_033001 NM_033003	0	1	0	4	0	7.34e-03
SPRR2A	Small Proline-RichProtein 2A	NM_005988	0	2	0	4	0	7.34e-03
S100A7	S100 Calcium-Binding Protein A7	NM_002963	0	1	0	3	0	2.03e-02
BAT1	HLA-B-Associated Transcript 1 A member of RNA helicases	NM_004640	0	1	0	3	0	2.03e-02

이 조금 떨어지지만 반대로 편평상피암종에서 발현이 증가된 유전자는 RPL37A, ACTB, 그리고 GTF2I 등 세 가지가 있었다.

유의성을 검증하는 또 한가지 방법은 더 많은 library를 분석하는 것이다. 그래서 이번에는 구강 내 정상 편평상피 (NCI_CGAP_HN7, 9, 11)와 침습성 암종 (NCI_CGAP_HN8, 12, 16)을 0과 0.05의 변수를 사용하여 비교해 보았다. 그 결과 4개의 EST와 KRT4, KRT13, RPS27A, GTT1, 그리고 S100A8 등이 암종에 비해 정상 편평상피에서 높은 발현을 보였고, RPL37A, ACTB, GTF2I, SPRR2A, S100A7, 그리고 BAT1 등은 암종에서 높은 발현을 보였다 (표 6).

한편, 유의성은 다소 떨어지지만 (p<0.0582) 앞의 cDNA xProfiler 분석에서 암종에만 발현되는 유전자로 분류되었던 HMG1과 FTL은 여전히 두 개의 암종 library에서 발견되었으나 정상 편평상피 library에선 전혀 발견되지 않았다.

표 5와 6의 분석 결과를 종합해 보면 정상 상피가 암종으로 전환하면서 KRT4, KRT5, KRT13, 그리고 S100A8과 같이 상피세포 특성과 관련된 유전자 발현을 상실하는 반면, 활발한 유전자 발현과 단백질 합성 (GTF2I, BAT1, RPL37A), 세포 증식 (FTL)과 침윤(HMG1)에 관여하는 유전자 발현이 증가하는 것을 알 수 있다. 한편, Northern blotting이나 RT-PCR같이 유전자 발현을 비교하는 실험에서 시료간의 RNA 양을 표준화하는데 종종 이용하는 beta-actin의 발현이 암종에서 증가된 것은 관찰된 차이가 단순한 실험상의 오차인지 아니면 진정한 차이인지 의문을 제기하였다.

이에 다른 유전자 발현 데이터에서도 beta-actin의 발현이 변화하는지 검색해보기로 했다. GEO 싸이트

표 7. 혀의 편평세포암종과 관련된 염색체 변이 (Unbalanced)

Band	Abnormality	Morphology	Site	Cases
1p11	add(1)(p11)	Squamous Cell carcinoma	Tongue	2
1p13	del(1)(p13)	Squamous Cell carcinoma	Tongue	4
1p22	del(1)(p22)	Squamous Cell carcinoma	Tongue	2
1q10	i(1)(q10)	Squamous Cell carcinoma	Tongue	3
1q11	add(1)(q11)	Squamous Cell carcinoma	Tongue	4
1q11	del(1)(q11)	Squamous Cell carcinoma	Tongue	3
3p11	add(3)(p11)	Squamous Cell carcinoma	Tongue	2
3p11	del(3)(p11)	Squamous Cell carcinoma	Tongue	3
3q10	i(3)(q10)	Squamous Cell carcinoma	Tongue	5
4p11	add(4)(p11)	Squamous Cell carcinoma	Tongue	2
5p10	i(5)(p10)	Squamous Cell carcinoma	Tongue	4
6q21	del(6)(q21q25)	Squamous Cell carcinoma	Tongue	2
6q25	del(6)(q21q25)	Squamous Cell carcinoma	Tongue	2
7q22	del(7)(q22q32)	Squamous Cell carcinoma	Tongue	2
7q32	del(7)(q22q32)	Squamous Cell carcinoma	Tongue	2
8p11	add(8)(p11)	Squamous Cell carcinoma	Tongue	6
8p21	del(8)(p21)	Squamous Cell carcinoma	Tongue	2
8q10	i(8)(q10)	Squamous Cell carcinoma	Tongue	11
9p13	del(9)(p13)	Squamous Cell carcinoma	Tongue	3
9q11	i(9)(q11)	Squamous Cell carcinoma	Tongue	2
10q10	i(10)(q10)	Squamous Cell carcinoma	Tongue	2
10q26	add(10)(q26)	Squamous Cell carcinoma	Tongue	2
11p11	add(11)(p11)	Squamous Cell carcinoma	Tongue	2
12p11	add(12)(p11)	Squamous Cell carcinoma	Tongue	3
12q24	add(12)(q24)	Squamous Cell carcinoma	Tongue	3
13p11	add(13)(p11)	Squamous Cell carcinoma	Tongue	3
14p11	add(14)(p11)	Squamous Cell carcinoma	Tongue	6
14q10	der(14;15)(q10;q10)	Squamous Cell carcinoma	Tongue	2
15p11	add(15)(p11)	Squamous Cell carcinoma	Tongue	8
15q10	der(14;15)(q10;q10)	Squamous Cell carcinoma	Tongue	2
19q13	add(19)(q13)	Squamous Cell carcinoma	Tongue	3
20q13	del(20)(q13)	Squamous Cell carcinoma	Tongue	2
21q10	i(21)(q10)	Squamous Cell carcinoma	Tongue	3
21q21	del(21)(q21)	Squamous Cell carcinoma	Tongue	2
Xq10	i(X)(q10)	Squamous Cell carcinoma	Tongue	2

에서 cDNA array를 이용해 37건의 신세포암종(renal cell carcinoma)을 정상 신상피조직과 비교해 암종의 발생에 의해 유전자 발현이 어떻게 변화하는지 분석한 결과를 찾았다[10]. 그들이 조사한 31,500 개의 cDNA 중 1,738 개가 암종에서 발현이 증가 또는 감소하였는데, beta-actin도 암종에서 증가된 유전자 중의 하나였다. 따라서 DGED 분석 결과가 진정한 발현 차이일 가능성을 높여주었다.

뿐만 아니라 1,738 개 유전자 목록은 RPL37A, Transferrin receptor, HMG1, GTF2I, 그리고 FTL도 포함하고있는데, FTL을 제외하고 나머지 유전자는 모두 신세포암종에서 증가한 것을 확인할 수 있었다[11].

4) 염색체의 Recurrent Aberration 검색

다음은 Mitelman 데이터베이스에서 암종에서 관찰

표 8. 유용한 데이터베이스 웹싸이트

Name	Web Address and Description
Human Genome Project Working Draft	bio-mirror.kr.apan.net/human_genome/
FISH-mapped BACs	cgap.nci.nih.gov/Chromosome/CCAP_BAC_Clones A set of BAC clones that have been mapped cytogenetically by FISH and physically by STSs to the human genome.
SNP Maps	www.ncbi.nlm.nih.gov/SNP/Database of SNPs and other genetic variations.
CGAP	cgap.nci.nih.gov/The Cancer Genome Anatomy Project
SAGE	www.ncbi.nlm.nih.gov/SAGE/Gene expression results from SAGE tags.
Gene Expression Omnibus	www.ncbi.nlm.nih.gov/geo/A public repository for expression data.
UniGene	www.ncbi.nlm.nih.gov/UniGene/Hs.Home Organization of transcribed sequences into gene-based clusters.
HomoloGene	www.ncbi.nlm.nih.gov/HomoloGene/ Putative homologies among human, mouse, rat, and zebrafish.
Homology Map	www.ncbi.nlm.nih.gov/Homology/Blocks of conserved synteny between mouse and human.
LocusLink	www.ncbi.nlm.nih.gov/LocusLink/Focal point for genes and associated information.
OMIM	www.ncbi.nlm.nih.gov/OMIM Guide to genes and inherited disorders.
GeneCards	nciarray.nci.nih.gov/cards/A database of human genes, their products and their involvement in diseases which offers concise information about the functions of all human genes.
MGC	mgc.nci.nih.gov/ The Mammalian Gene Collection to provide a complete set of full-length (open reading frame) sequences and cDNA clones of expressed genes for human and mouse.
Mitelman Database	cgap.nci.nih.gov/Chromosomes/Mitelman A Database of Chromosome Aberrations in Cancer.
BLAST the human genome	www.ncbi.nlm.nih.gov/genome/seq/HsBlast Compare your sequence to the genome or its gene products.
VAST search	www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch A structure-structure similarity search service.

된 염색체 변이를 검색해 보았다. Recurrent Chromosome Aberration 검색엔진은 염색체 (all 또는 특정 염색체), Arm (Both, p, 또는 q), Band (all 또는 특정 band), Aberration type (all, balanced, 또는 unbalanced), 부위 (all 또는 62가지 조직 중 하나), 종양 (all 또는 199가지 종양 중 하나) 등의 변수를 선택한 후 검색할 수 있다. 부위 중 혀를 그리고 종양 중 편평세포암종을 선택하여, 혀에 발생한 편평세포암종과 관련된 모든 염색체 변이를 검색하였다.

그 결과 balanced abberation으로 t(8:20)(q11:p11)이 2건 보고된 바 있었고, 표 7과 같이 35가지 종류의 unbalanced aberration이 검색되었다. 혀 및 구강내 편평세포암종에서 발현이 감소 또는 증가된 유전자 목록 (표 6)에서 유전자의 위치가 염색체 변이부위와 일치하는 것이 있는지 찾아보았다. FTL이 염색체 19q13.3에 위치하는데, 3건의 add(19)(q13)이 보고된 바 있었다.

또한 부위에 혀 대신 구강을 넣어 검색하자 38종류의 unbalanced aberration이 검색되었는데, 3건의 add(19)(q13)과 SPRR2A와 S100A7 유전자가 위치하는 add(1)(q21) 2건이 눈에 띄었다.

이상의 데이터베이스 분석결과는 다음과 같은 연구에 이용될 수 있다. 먼저 표 5와 6에 있는 EST와 유전자의 발현 차이가 실제로 관찰되는지 정상과 암종 조직표본에서 in situ hybridization을 사용해 검증

해 볼 필요가 있다. 실험으로 검증된 유전자들은 그 유전자의 발현 양상이 암종의 분류나 예후와 어떤 관련이 있는지 보다 면밀한 연구를 할 수 있다. 또한, SNPs 지도 검색 결과 KRT4, KRT13, GTT1, RPL37A, GTF2I, SPRR2A, 그리고 BAT1에는 SNP가 존재하므로, 이들 유전자 다형성이 암종 발생과 어떤 관련이 있는지, 각 유전자가 만든 단백질이 암종 발생과 관련해 어떻게 기능 하는지 연구할 수 있다. EST의 경우, 발표된 게놈 데이터베이스에 blast search (www.ncbi.nlm.nih.gov/genome/seq/HsBlast)를 함으로써, EST가 진정한 유전자인지 확인하고, 클론을 구입하여 온전한 유전자를 클론할 수 있다. 이렇게 새로 밝혀진 유전자가 암종의 발생에 어떤 역할을 하는지 중요한 연구를 할 수 있을 것이다.

마지막으로 지금까지 언급한 데이터베이스와 기타 유용한 웹싸이트를 표 8에 정리하였다.

참 고 문 헌

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860-921. 2001.
2. http://genome.cse.ucsc.edu
3. Venter, J. C. et al. The sequence of the human genome. Science 291:1304-51. 2001
4. The BAC Resource consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature 409, 953-958 (2001)
5. The International SNP MAP Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409, 928-933. 2001.
6. http://cgap.nci.nih.gov/
7. http://cgap.nci.nih.gov/Reagents
8. Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W. Serial Analysis of Gene Expression. Science 270 : 484-487. 1995.
9. Dillon, D. A., D'Aquila, T., Reynolds, A. B., Fearon, E. R., Rimm, D. L. The expression of p120ctn protein in breast cancer is independent of alpha- and beta-catenin and E-cadherin. Am. J. Path. 152 : 75-82, 1998.
10. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc =GSE3
11. http://www.dkfz-heidelberg.de/abt0840/whuber/rcc/