마이크로어레이 자료의 분석

Analysis of microarray data

질병관리본부 국립보건연구원 유전체센터 바이오과학정보과
조성범

Ⅰ. 들어가는 말

마이크로어레이는 기능 유전체학 연구에 있어서 혁명적인 발상의 전환을 가져온 기술이다. 기존의 방법이 개별 유전자 발현을 측정하는 반면, 마이크로어레이 기술은 한 번의 실험으로 수천에서 수만 개의 유전자 발현을 측정할 수 있다. Shena 등은 1995년에 실험실에서 직접 제작한 마이크로어레이 기판을 이용하여 세포에서 발현하는 전사체(transcriptome)의 발현을 측정하였다[1]. 이러한 유전자 발현 마이크로어레이(gene expression microarray)의 등장으로 한 세포에서 대용량의 유전체정보를 손쉽게 측정할 수 있는 고효율실험기법(high-throughput technology)이 발달하기 시작하였다. 그 결과 단일염기다형성(single nucleotide polymorphism, SNP), 복제수변이(copy number variation, CNV), 염색질 면역침강법(chromatin immunoprecipitation, ChIP)등에 대한 정보를 마이크로어레이 실험으로 발굴할 수 있는 기술들이 개발되었다.

마이크로어레이 기술이 발달하면서 대용량 유전체자료에 대한 분석이 유전체 및 생물정보학 연구에서 중요한 주제로 자리 잡기 시작하였다. 변수의 종류가 상대적으로 제한적이고 시료의 수가 많던 기존의 자료 형태에서 변수의 양이 기하급수적으로 늘어난 고차원자료(high-dimensional data)의 형태로 바뀌면서 기존의 통계적인 분석방법론을 곧바로 적용하는 것이 어렵게 되었다. 그 결과, 마이크로어레이 자료에 적합한 고유의 방법론들이 계속해서 등장하였다. 이 글에서는 현재까지 마이크로어레이자료의 분석에서 주로 사용하고 있는 분석방법 중에서 마이크로어레이 자료의 실험 편차(experimental bias)를 보정하는 정규화(normalization) 방법에 대한 소개는 제외하고, 정규화 이후에 생물학적 의미를 발견하기 위한 분석법에 대하여 요약하여 소개하고자 한다.

Ⅱ. 몸 말

1. 차별발현 유전자(differentially expressed gene, DEG) 분석법

마이크로어레이 자료를 이용한 분석 중에서 가장 기본적인 것이라 할 수 있는 것이 차별발현유전자의 분석이다. 이 분석은 동일한 유전자의 평균 발현량이 서로 다른 조건에서 유의하게 다른지를 분석하는 방법론이다.

가장 많이 사용하는 방법론 중의 하나는 SAM(significant analysis of microarray) 방법으로, 기존의 T-test를 변형시킨 분석방법으로 variance shrinkage 방법과 bootstrapping 기법을 적용하여 마이크로어레이 자료의 특성에 맞는 분석을 시행한다[2]. 서로 다른 두 군에 대한 계산뿐만 아니라 여러 개의 군집으로 나누어 있는 경우에도 적용가능하고 생존 정보가 있는 자료의 분석도 가능하다. 최근에는 마이크로어레이 실험의 대안으로 많이 시행하고 있는 RNA-seq의 분석도 적용할 수 있도록 개발되었다 (Figure 1).

SAM과 함께 많이 사용되고 있는 방법은 Cyber-T 방법이다. 이 방법 역시 T test를 기반으로 하지만 베이지안(Bayesian) 통계 기법을 사용하여 서로 다른 군의 분산을 추정하는 특징을 가지고 있다[3]. 그리고 이웃하는 유전자의 분산을 고려하여 유전자 발현량의 분산을 추정하고 있다. 특히, 이 방법은 샘플 수가 적은 자료에서 탁월한 성능을 발휘하는 것으로 알려졌다.

†

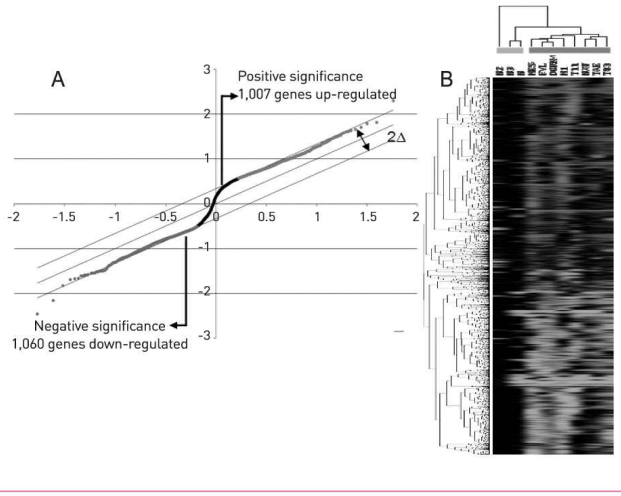

Figure 1. SAM analysis results[2]

2. 차별공발현 유전자쌍(differentially coexpressed gene pairs) 분석법

차별발현 유전자가 서로 다른 조건에서 평균적인 발현량이 유의하게 변하는 유전자를 찾아내는 방법이라면, 차별공발현 유전자쌍 분석법은 서로 다른 조건에서 두 유전자의 공발현 (coexpression)이 유의하게 변화하는 양상을 분석하는 방법이다[4](Figure 2).

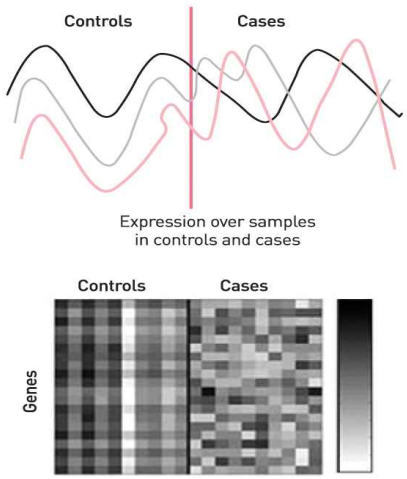

Figure 2. Conceptual diagram of differentially coexpression[9]

유전자쌍의 공발현의 계산은 두 유전자 발현의 연관정도(correlation)로 파악하는 것이 보통이다. 이 때 Pearson의 상관계수를 가장 많이 사용하고 그 외에도 Spearman, Kendall의 상관계수 및 entropy를 이용한 공발현 계산방법도 사용한다. 차별 공발현쌍의 계산은 서로 다른 조건에서 구한 공발현 정도가 유의하게 변하는 지에 대한 검정을 시행하여 수행한다.

차별발현 유전자와 차별공발현 유전자쌍에 대한 차이는 Figure 3에 잘 나타나있다. 차별발현 유전자는 한 조건 안에서의 평균 발현량만을 고려하고, 차별공발현 유전자쌍의 경우 한 조건 안에서는 잘 정렬된 발현양상(expression pattern)을 보이던 유전자쌍의 집합들이 다른 조건(질병 등)에서는 흩어진 발현양상을 고려한다. 이러한 차별공발현 유전자쌍들은 질병 기전이나 생물학적인 현상의 발현에 관여하는 것

으로 잘 알려졌다.

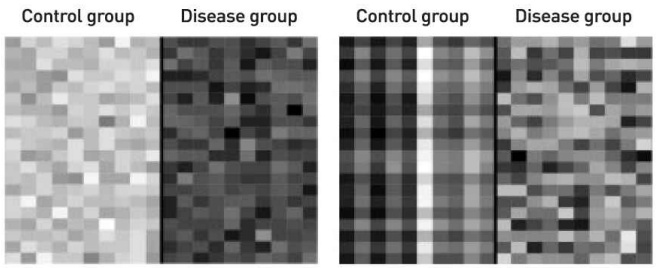

Figure 3. Differentially expressed genes vs. differentially coexpressed gene pairs[9]

3. 기능적 주석 (functional annotation)

유전자 발현 마이크로어레이자료를 분석한 후에 가장 먼저 접하게 되는 문제는 분석결과에 대한 해석이다. 마이크로어레이자료가 수 만개의 유전자에 대한 정보를 포함하고 있기 때문에 다중검정교정(multiple testing correction)을 시행하고 난 후에도 수십에서 수백 개의 유전자가 결과에 나타날 수 있다. 이런 경우에는 각 유전자에 대한 주석을 통한 해석은 많은 시간을 요구하기 때문에 어려움이 있다. 이 문제를 해결하기 위해서 기능적 주석방법이 고안되었다. 기능적 주석방법은 유전자군에서 유의하게 반복적으로 나타나는 생물학적인 주제어의 검출을 수행하고 있는 방법론으로 기능적 농축 검정(enrichment test)으로 불리기도 한다[5].

기능적 주석 혹은 농축 검정의 원리는 한 무리의 유전자들안에 특정 생물학적인 범주에 해당하는 유전자가 얼마나 빈번하게 나타나는 지에 대한 검정방법이다. 예를 들어 차별발현된 100개의 유전자군 안에 80개의 유전자가 p53 pathway에 속하는 유전자라고 하면 p53 pathway가 서로 다른 차별발현되는 기전과 연관이 있다고 추측할 수 있다. 이 때 만약 p53 pathway에 속하는 유전자들의 수가 많다고 가정하면, 100개의 유전자 중에서 80개의 p53 pathway 관련 유전자는 우연히 결과에 들어가 있을 수도 있다. 이러한 가능성을 검정하기 위해서 농축 검정(enrichment test)을 시행하고, 주로 2×2 표 기반의 Chi-square 검정이나 Fisher's exact test를 이용하여 검정을 수행한다. 이 때 전체 유전자는 마이크로어레이 플랫폼에 심어진 유전자로 하거나, 알려진 모든 유전자를 모집단으로 설정할 수 있다.

농축 검정의 과정은 다음과 같다. 먼저, 유의한 결과에 있는 유전자군과 그렇지 않은 유전자군으로 전체 유전자를 분류한다. 그리고 특정 생물학적 범주(biological pathway 혹은 gene ontology, GO 등)에 속하는 유전자들이 각 군에 몇 개씩 있는 지를 파악하고 2×2 표를 완성하여 통계적인 검정을 수행한다. 이 과정을 각각의 생물학적 범주에 따라서 반복한 후에 다중검정교정을 통하여 교정된 p값에 따라서 유의한 결과를 정한다. 유의한 결과를 보이는 pathway나 GO는 서로 다른 조건에서 발생하는 생물학적 현상의 기전에 대한 단서를 제공할 수 있다.

4. 군집분석(clustering analysis)

군집분석은 서로 다른 요소 간의 유사성을 바탕으로 서로 유사한 성질을 지니는 요소끼리 군집을 형성하는 분석방법을 말한다. 이 방법은 요소의 수가 많을 때 그 요소들의 특성과 요소간의 관계를 파악할 때 많이 사용하는 다변량 통계기법이다.

마이크로어레이 자료에서 군집분석은 하나의 유전자 단위에서 수행하는 분석과는 다르게 각 유전자들을 그 발현양상에 따라서 여러 개의 군집으로 나누는 분석방법을 말한다. 마이크로어레이의 군집분석은 다양한 의미로 해석될 수 있는데, 많은 연구자들이 발현모듈(expression module)의 관점에서 군집분석의 결과를 해석한다.

발현모듈이란 동일한 전사조절인자의 영향으로 유사한 발현양상을 보이는 유전자들의 집합을 말한다. Figure 4는 효모의 유전자 발현 마이크로어레이 자료의 군집분석 결과이다[6]. 그림에서 보이는 것과 같이 효모에 서로 다른 조건을 주었을 때 발현양상이 달라지는 것을 관찰할 수 있다. 이 때 군집분석을 통해서 유사한 발현양상을 보이는 유전자들의 군집을 파악할 수 있다. 한 군집 내에 속하는 유전자들은 그 조건에 반응하는 효모의 생물학적 과정에서 유사한 기능을 하는 유전자들일 가능성이 높다는 것을 GO 분석을 통해서 밝히고 있다. 군집분석을 시행한 후에 각 군집이 가지는 생물학적인 의미는 위에서 설명한 기능적 주석 방법으로 설명할 수 있다.

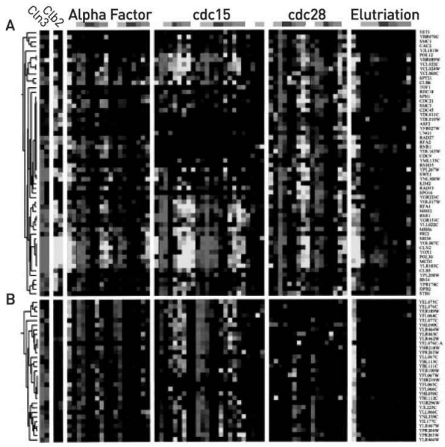

Figure 4. Change of yeast gene expression patterns in different conditions[6]

5. 분류분석(classification analysis)

분류분석 방법은 군집분석 방법과 마찬가지로 기존에 많이 사용하던 다변량 통계방법이다. 이 분석법은 서로 다른 군을 가지는 한 집단에서 주어진 변수의 특성을 파악하여 서로 다른 군에서 이 특성들이 변하는 정도에 관한 규칙을 수식으로 정리한다. 그리고 어느 군인지를 모르는 집단이 있을 때 이 집단의 구성원들이 어느 군에 속하는 지를 판단 혹은 예측할 수 있는 방법을 말한다.

마이크로어레이 자료에서 분류분석은 주로 생물학적 표지자(biomarker)의 발굴에 쓰이고 있다. 특히 암의 마이크로어레이 분석에 많이 사용되었는데, Golub 등은 유전자 발현 마이크로어레이 자료를 이용하여 급성 골수성 백혈병과 급성 림프구성 백혈병을 임상 정보 없이 정확하게 분류하는 연구 결과를 발표하였다[7](Figure 5). Van't Veer 등은 유방암의 마이크로어레이 자료를 이용하여 유방암 환자의 예후를 예측하는 70개의 유전자군을 발견하였고 이를 바탕으로 MamaPrint라는 진단 상품이 개발되었다[8].

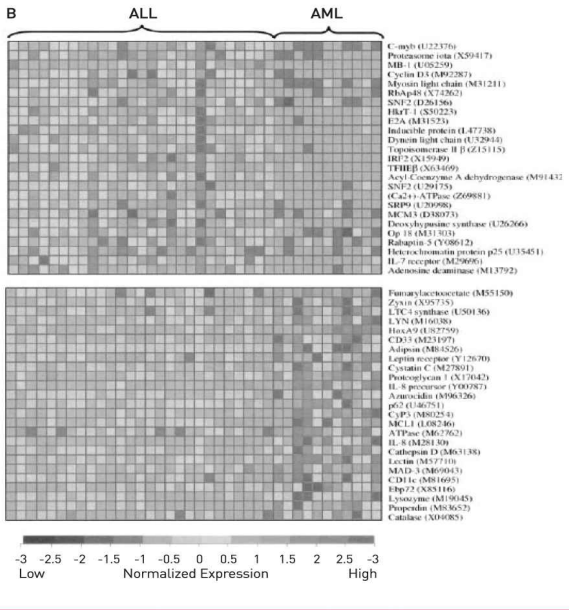

Figure 5. Gens classifying acute myelogenous and lymphoblastic leukemia[7]

III. 맺는 말

지금까지 마이크로어레이 자료의 분석에 자주 사용되고 있는 방법에 대하여 살펴보았다. 이러한 방법들은 마이크로어레이 자료가 등장한 초기에 확립되어 지금까지 활발히 사용되고 있다. 이 방법들은 분석 목적에 따라서 변형되어 사용되기도 하고 전혀 다른 방법들이 사용되기도 한다. 그리고 최근에는 서열 기반의 RNA-seq 자료도 등장하여 마이크로어레이 기법과 병행하여 사용되고 있다. 그러나 서로 다른 실험 기법을 사용하더라도 기능 유전체학의 연구 목적으로 위의 방법들은 동일하게 적용될 수 있다. 현재의 추세로 볼 때 유전체 기능 연구를 위하여 마이크로어레이 및 RNA-seq등의 전사체 자료의 활용이 더욱 더 중요해지고 있다.

IV. 참고문헌

1. Schena, M., Shalon, D., Davis, R.W. and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995 270;467-470.
2. Tusher VG, Tibshirani R, Chu G.Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001 24;98(9):5116-21.
3. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics. 2001 Jun;17(6):509-19.
4. Lai Y. Wu B, Chen L, Zhao H A statistical method for identifying differential gene-gene co-expression network. Bioinformatics. 2004 22;20(17):3146-3155.
5. Huang DW, Lempicki RA(2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.". Nature Protocols 2008 4(1): 44-57.
6. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces

cerevisiae by microarray hybridization. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Mol Biol Cell. 1998 9(12):3273-97.
7. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Science. 1999 286(5439):531-7.
8. Gene expression profiling predicts clinical outcome of breast cancer. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Nature. 2002 415(6871):530-6.
9. Finding disease specific alterations in the co-expression of genes Dennis Kostka and Rainer Spang. Bioinformatics. 2004 20.i194-i199.