# 프마이크로어레이 데이터의 유전자 집합 및 대사 경로 분석
## (Gene Set and Pathway Analysis of Microarray Data)

김 선 영
한국생명공학연구원 의약유전체 연구센터

Seon-Young Kim, Ph. D.
*Functional Genomics Research Center, Division of Molecular Therapeutics, Korea Research Institute of Bioscience and Biotechnology, 52 Eoeun-dong, Yuseong-gu, Daejeon 305-333, Korea.*
*kimsy@kribb.re.kr.*

## Abstract

Gene set analysis is a new concept and method to analyze and interpret microarray gene expression data and tries to extract biological meaning from gene expression data at gene set level rather than at gene level. Compared with methods which select a few tens or hundreds of genes before gene ontology and pathway analysis, gene set analysis identifies important gene ontology terms and pathways more consistently and performs well even in gene expression data sets with minimal or moderate gene expression changes. Moreover, gene set analysis is useful for comparing multiple gene expression data sets dealing with similar biological questions. This review briefly summarizes the rationale behind the gene set analysis and introduces several algorithms and tools now available for gene set analysis.

## I. 서 론

1995년 처음으로 도입된 이후 microarray 방법은 생물학 연구에 큰 영향을 끼쳐 왔으며, 이제는 거의 모든 실험실에서 일상적으로 사용하는 도구가 되었다. 초창기 이래 microarray 칩 제작 기술은 끊임없이 발전해 오고 있는데, 더욱 정확하고 정보량이 많은 고밀도 칩들이 개발되어 널이 쓰이고 있다. 더불어 microarray에서 생산되는 정보를 분석하고 해석하는 방법들 또한 꾸준히 발전하여 실험자들이 생산한 방대한 양의 자료를 해석하는 데 많은 도움을 주고 있다.

Microarray 데이터 분석은 스캐너로 읽은 형광 이미지로부터 각 유전자의 발 현양을 계산하는 데서 출발한다. Single channel platform인 Affymetrix 회사의 microarray data와 dual channel platform인 cDNA 혹은 spotted oligonucleotide microarray data 각각에 대해서, 형광 이미지로부터 유전자의 발현량을 구하는 다양한 방법들이 개발되어 쓰이고 있다. 형광 이미지에서 유전자의 발현량을 계산하는 과정은 이미지 상에서 부정확한 부분을 제거하고, 각 시료 간의 차이를 보정하는 표준화(normalization) 과정을 포함한다. 이러한 데이터 전 처리(preprocessing) 과정을 거친 후 연구자들은 비로소 생물학적 의미를 찾아내는 분석을 하게 된다.

현재 microarray 데이터를 사용하는 연구자들은 크게 두 가지 방법으로 데이터를 분석한다. 첫 번째 방법은 Eisen 그룹에 의해 도입되어 널리 쓰이고 있는 clustering 방법으로, 각 유전자들의 발현값들 사이의 유사도(similarity)를 계산하여, 비슷한 유전자 발현

양상을 보이는 샘플 (samples)과 유전자 (genes)들을 찾아내는 방법이다.[1] 두 번째 방법은 통계적인 방법 (예컨대, t-test나 ANOVA)을 적용하여 전체 데이터로부터 일정 기준을 통과하는 수십 혹은 수백 개의 유전자들을 선별하고, 이들 선별된 유전자들로부터 생물학적 의미를 찾아내는 방법이다. 두 번째 방법에 대해 좀더 덧붙이면, 초창기의 연구자들은 그들의 생물학적 지식을 동원하여 선별된 유전자들을 여러 그룹들로 구분, 정리하여 논문에 기술하였는데, 이제는 이러한 과정을 자동적으로 수행하는 수십 가지 분석 도구들이 개발되어 널리 쓰이고 있다. 이들 분석 도구들은 대부분 gene ontology(GO) 정보와 pathway 정보를 바탕으로 선별한 수십, 수백의 유전자들로부터 의미 있게 많이 나타나는(over-represented) gene ontology 정보 혹은 pathway 정보를 찾아낸다. 최근에 Khatri[2] 등과 Curtis[3] 등은 이들 분석 도구들을 체계적으로 비교, 정리한 바 있다.

## II. 기존 방법의 문제점 및 유전자 집합(gene set) 분석

기존의 gene ontology 혹은 pathway 분석 방법들은 거의 대부분 hypergeometric distribution(비복원 추출, sampling without replacement, 에 의해 만들어 지는 분포)을 바탕으로, Fisher's exact test나 $\lambda^2$-test를 사용하여 각 유전자 집합의 통계적인 유의성을 추론한다. 이러한 방법들은 over-representation analysis (ORA)라 불리는데, 다음에 기술하는 몇 가지 단 점을 가지고 있다.[4,5] 첫째, 전체 유전자 집단으로부터 일부 유전자들을 선별하는 기준 자체가 임의적이다. 대개의 경우 2 배 이상 혹은 이하로 변하고, 통계 검사 (가령, *t*-test나 ANOVA)의 유의성이 0.05 이하라는 두 가지 기준을 사용하는 데, 이들 두 가지 기준은 실험자가 편의상 임의로 설정한 기준일 뿐 실제 생물체 혹은 세포 내에서 일어나는 생물학적 변화를 구분하는 기준은 아니다. 두 번째, 기존의 gene ontology나 pathway 분석 도구들은 유전자 이름만을 입력 정보로 사용할 뿐 발현량 변화 정도나 그 통계적 유의성 등은 사용하지 않으므로, 많은 양의 유용한 정보를 활용하지 않은 채 버린다. 끝으로, 일정 기준을 통과하는 유전자들의 수가 아주 적거나, 아예 없는 경우, 기존의 방법으로는 gene ontology 혹은 pathway 분석 자체를 할 수 없다.

몇몇 연구자들은 기존 ORA(over-representation analysis) 분석 방법의 단점을 극복하고자 다른 접근 방법을 시도하여 왔다. Functional class scoring (FCS) 혹은 유전자 집합 분석 (gene set analysis)이라고 명명된 이 방법은, 수십 혹은 수백 개의 유전자를 선별한 후 분석하는 대신, 모든 유전자를 다 사용하면서 그 중 의미 있는 유전자 집합을 골라낸다(그림 1). FCS 방법은 2002년 Pavlidis 및 동료 연구자들이 처음으로 시도하였는데, ORA에 비해 더욱 민감하게 의미 있는 유전자 집합을 찾아낼 수 있음이 보고되었다.[4] 그러나 FCS 방법이 주목을 받게 된 것은, 2003년 Mootha외의 연구자들이 당뇨병 환자와 정상인의 근육 조직 간의 유전자 발현 차이를 분석하는 논문을 발표하면서부터이다.[6] 이들은 처음의 분석에서 당뇨병 환자 근육과 정상 근육 조직 사이에 현저하게 차이 나는 유전자들을 골라낼 수 없었는데, 149개의 유전자 집합(gene set)을 정의하고, 이들 유전자 집합 수준에서 두 그룹 간의 차이를 조사한 결과, 당뇨병 환자들의 근육 조직에서 산화적 인산화 과정(oxidative phosphorylation)에 관여하는 유

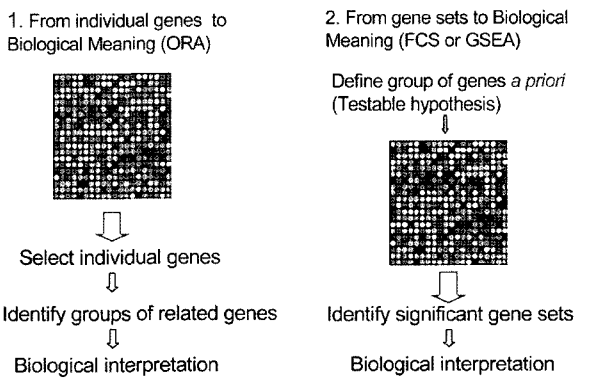


**그림 1.** Two approaches in microarray analysis and interpretation

표 1. Comparison of Gene Set Analysis Softwares

| | PAGE | GSEA | ErmineJ | MEGO | Catmap | T-profiler |
|---|---|---|---|---|---|---|
| Used Statistics | Fold Change | Rank | P-value | Fold change | P-value | T-statistic |
| Statistical test | *z*-test | permutation (sample) | permutation (gene) | | permutation (sample, gene) | *t*-test |
| Speed | Fast | Slow | Moderate | Fast | Slow | Fast |
| Standalone software | YES | YES | YES | YES | YES | NO |
| GUI | NO | YES | YES | YES | NO | NO |
| Web server | YES | NO | NO | NO | NO | NO |
| Organism | H, M, R, Y | H, M | H, M, R | H | H | Y |
| Gene Sets | GO Pathways Chromosomes | GO Pathways Chromosomes | GO Pathways Chromosomes | GO Pathways Chromosomes | GO Pathways Chromosomes | GO Pathways Chromosomes |

전자들의 발현이 적은 폭이지만 일관되게 변화하는 것을 밝혀냈다. 이들은 그들의 분석 방법을 gene set enrichment analysis(GSEA)라 명명하였다.

Gene Set Enrichment Analysis(GSEA)가 발표된 이래, 같은 개념을 사용하는 많은 다양한 앨고리즘과 분석 도구들이 개발되어 왔는데, 이들 도구들은 2003년에 발표되었던 GSEA 방법이 가지고 있던 여러 가지 문제점들을 개선해 왔고, 사용자가 편리하게 사용할 수 있는 프로그램 환경을 제공해 왔다.[7-12] 표 1은 이들 여러 앨고리즘과 프로그램들을 여러 측면에서 비교하고 있다. 이들은 프로그램에 입력할 때 사용하는 입력값(input value)에서, 사용하는 통계 모형 (statistical model), 사용자 환경(user interface) 및 사용 가능한 microarray platform 등에서 서로 차이가 난다. 그 중 가장 중요한 측면으로 사용하는 통계 모형에 대해 좀더 상세히 언급하면, 각 유전자 집합(gene set)의 통계적 유의성을 추론하기 위한 배경 분포(background distribution)를 구하는 방법은 유전자 순열(gene permutation)과 시료 순열(sample permutation) 이렇게 두 가지로 크게 구분된다. 시료 순열(sample permutation)은 microarray data set 내의 시료(sample)들을 임의로 바꾸어 만든(보통 1,000번 수행) 데이터 집합들로부터 구한 값들을 배경 분포로 사용하는데, 정확한 통계 분석이 가능한 반면, 일정 수 이상의 시료(적어도 8개 이상)가 있어야만 의미 있는 유전자 집합(gene set)을 찾아낼 수 있고, 또 시간이 많이 걸리는 단점을 가지고 있다. 반면, 유전자 순열(gene permutation) 방법은 빠르고, 시료 수의 크기에 제한을 받지 않는 반면, 한 유전자 집합(gene set) 내에 있는 유전자들의 상당수가 correlation을 보이는 경우, 실제로는 의미가 있지 않은 유전자 집합(gene set)을 의미 있게 선별하는 오류를 범할 가능성이 있다.[5,8-10]

유전자 집합 분석(gene set analysis)에서 사용하는 앨고리즘만큼 중요한 것은 유전자 집합(gene set) 데이터베이스다. 사용하는 유전자 집합(gene set)의 내용과 범위가 다양하면 다양할수록, 유전자 집합 분석은 많은 의미 있는 정보와 통찰을 제공할 것이다. 현재 유전자 집합 분석 (gene set analysis)을 제안한 여러 연구자들은 앨고리즘과 더불어 다양한 유전자 집합(gene set)들을 데이터베이스로 제공하고 있다. 그 중, 대표적인 것은 GSEA를 발표했던 연구 그룹에서 제공하는 Molecular Signature Database (MolSigDb, http://www.broad.mit.edu/gsea/msigdb/msigdb_index.html)인데, 1) chromosomal location, 2) biological pathways, 3) transcriptional regulatory elements, 4) coregulated gene sets 이렇게 네 가지 범주로 나누어서 다양한 유전자 집합 (gene set)을 제공하고 있다. 그리고 이러한 유전자 집합 데이터베이스는 여러 연구 결과가 쌓여가면서 더욱 다양하고 풍부해 질 것이다.

유전자 집합 분석(gene set analysis)은 한 gene expression 데이터에서 중요한 생물학적 의미와 pathway를 밝혀내는 데 유용할 뿐 아니라, 같은 주제를 다루는 서로 다른 데이터들을 비교, 검토하는 데

도 유용하다. Microarray를 이용한 연구에서 여러 연구자들이 초창기부터 경험했던 심각한 문제점들 중 하나는 서로 다른 실험실에서 생산된 microarray 데이터 분석 간에 공통점이 별로 없다는 점이었다. 대개의 경우 비교하고자 하는 논문들에 실린 수십, 수백 개의 선택된 유전자 리스트들 중에서 서로 공통된 유전자가 얼마나 있는가를 조사하는데, 불행하게도 많은 연구에서 공통된 유전자의 비율은 극히 낮았다. 이러한 심각한 불일치는 여러 가지 원인 - platform, 실제 시료, 분석 방법, 그리고 심지어 실험 기술의 차이-들 탓으로 돌려졌고, 이러한 문제점들을 극복하기 위한 노력들이 계속 이루어져 왔다.[7] 그런데, 같은 생물학적 질문을 연구하는 서로 다른 microarray 데이터들을 비교할 때, 유전자 집합 (gene set) 수준에서 두 데이터를 비교하면 대부분의 경우에서 개별 유전자들을 비교할 때 보다 훨씬 높은 유사성을 관찰할 수 있다(그림 2 참조).[7-8] 이는 서로 다른 데이터에서 관찰할 수 있는 생물학적 변화가 개별 유전자 수준보다는 유전자 집합(gene set) 수준에서 더 정확하게 파악됨을 뜻하는데, 개개 유전자의 발현 변화는 각 상황에 따라 달라질 수 있지만, 전체적인 생물학적 변화는 크게 차이가 없기 때문이다. 그러므로 유전자 집합 분석(gene set analysis)은 meta-analysis와 더불어 서로 다른 여러 microarray 데이터들로부터 공통된 혹은 차이 나는 생물학적 의미를 끌어내는 데 있어 중요한 분석 방법이 될 수 있다.

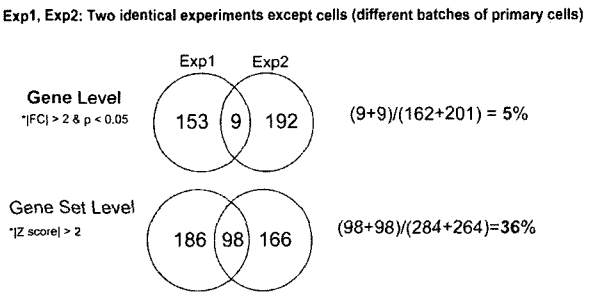


**그림 2.** Comparison of multiple data sets: Gene level vs. Gene Set level

## Ⅲ. 맺음말

본 글에서는 최근에 등장하여 중요한 도구로 여겨지는 유전자 집합 분석(gene set analysis)에 대해 간략하게 살펴보았다. 유전자 집합 분석 방법은 초기에는 독립된 소프트웨어로 등장했는데, 그 강력한 분석 능력으로 인해 이제 종합적인 microarray 분석 소프트웨어의 일부분에 포함되기 시작하고 있다. 예컨대, 국내 연구자들도 많이 사용하고 있는 BRB ArrayTools (http://linus.nci.nih.gov/~brb/download.htm)도 버전 3.4에 gene set enrichment analysis를 포함시켰고, 암 조직의 유전자 발현 데이터베이스로는 가장 방대한 Oncomine 데이터베이스(http://www.oncomine.org)도 gene set enrichment analysis를 포함시켰다. 이러한 경향은 다른 microarray 소프트웨어들에도 마찬가지로 나타날 것으로 여겨지며, 조만간 대부분의 microarray 소프트웨어들은 그 기본 기능 중의 하나로 유전자 집합 분석(gene set analysis)을 포함하게 될 것이다.

## Ⅳ. 참고 문헌

1. Eisen, M.B., *et al.*, *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. 95(25): p. 14863-8.
2. Khatri, P. and S. Draghici, *Ontological analysis of gene expression data: current tools, limitations, and open problems.* Bioinformatics, 2005. 21(18): p. 3587-95.
3. Curtis, R.K., M. Oresic, and A. Vidal-Puig, *Pathways to the analysis of microarray data.* Trends Biotechnol, 2005. 23(8): p. 429-35.
4. Pavlidis, P., D.P. Lewis, and W.S. Noble, *Exploring gene expression data with class scores.* Pac Symp Biocomput, 2002: p. 474-85.
5. Tian, L., *et al.*, *Discovering statistically significant pathways in expression profiling studies.* Proc Natl

Acad Sci U S A, 2005. 102(38): p. 13544-9.

6. Mootha, V.K., *et al.*, *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nat Genet, 2003. 34(3): p. 267-73.
7. Kim, S.Y. and D.J. Volsky, *PAGE: Parametric Analysis of Gene set Enrichment.* BMC Bioinformatics, 2005. 6(1): p. 144.
8. Subramanian, A., *et al.*, *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.
9. Lee, H.K., *et al.*, *ErmineJ: tool for functional analysis of gene expression data sets.* BMC Bioinformatics, 2005. 6: p. 269.
10. Breslin, T., P. Eden, and M. Krogh, *Comparing functional annotation analyses with Catmap.* BMC Bioinformatics, 2004. 5(1): p. 193.
11. Boorsma, A., *et al.*, *T-profiler: scoring the activity of predefined groups of genes using gene expression data.* Nucleic Acids Res, 2005. 33(Web Server issue): p. W592-5.
12. Tu, K., H. Yu, and M. Zhu, MEGO: *gene functional module expression based on gene ontology.* Biotechniques, 2005. 38(2): p. 277-83.