The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.2.241
31(2), 241–251

A study on alternatives to the permutation test in gene-set analysis

Sunho Lee[a,1]

[a]Division of Mathematics and Statistics, Sejong University

(Received January 23, 2018; Revised February 21, 2018; Accepted February 26, 2018)

Abstract

The analysis of gene sets in microarray has advantages in interpreting biological functions and increasing statistical powers. Many statistical methods have been proposed for detecting significant gene sets that show relations between genes and phenotypes, but there is no consensus about which is the best to perform gene sets analysis and permutation based tests are considered as standard tools. When many gene sets are tested simultaneously, a large number of random permutations are needed for multiple testing with a high computational cost. In this paper, several parametric approximations are considered as alternatives of the permutation distribution and the moment based gene set test has shown the best performance for providing p-values of the permutation test closely and quickly on a general framework.

Keywords: gene set analysis, moment, multiple testing, permutation test

1. 서론

질병의 진단이나 치료에 관여하는 바이오마커를 찾아내기 위한 마이크로어레이 자료 분석은 표본의 표현형(암 또는 정상, 돌연변이 발생 여부나 병기 등)에 따라 발현의 차이를 보이는 특이발현 유전자나 유전자집합을 추출하는데 목적이 있다. 이때 사용되는 유전자집합은 동일한 기능의 대사경로에 참여하였거나 염색체 위치가 같은 유전자들, 또는 전장유전체연관분석에서 밝혀진 특정 형질과 관련된 유전자들의 모임이며, 어떤 집합이 질병의 발생에 관련이 있는지에 대한 판단은 표현형에 따라 발현 형태에 유의한 차이가 있는지 검정한 결과에 따른다. 이 때 검정은 비교대상을 어떻게 정의하는가에 따라 두 가지의 귀무가설이 가능하다 (Tian 등, 2005). 경쟁귀무가설(competitive null hypothesis)은 유전자집합과 표현형 사이의 연관성을 비슷한 크기의 다른 유전자집합과 비교하고 자립귀무가설(self-contained null hypothesis)은 집합에 속한 유전자들로부터 만들어 낼 수 있는 우연한 관계의 연관성들과 비교하는 것이다.

유전자집합을 분석하는 방법은 Mootha 등 (2003)과 Subramanian 등 (2005)이 제안한 gene set enrichment analysis (GSEA)에서 비롯하여, 중심극한정리를 이용한 Kim과 Volsky (2005)의 모수적 방법, 유전자들 사이의 상관관계를 이용한 Rahmatallah 등 (2014) 등 많은 방법이 있다. Ackermann과 Strimmer (2009)은 여러 방법들을 비교하였고 Mooney과 Wilmot (2015)는 유전자집합 분석과정을 체계적으로 설명하였다.

[1]Professor, Division of Mathematics and Statistics, Sejong University, 209 Neungdongro, Kwangjinku, Seoul 05006, Korea. E-mail: leesh@sejong.ac.kr

유의한 유전자집합을 찾는 방법은 검정통계량이 유도된 배경에 따라 특이발현 정도가 큰 유전자들이 포함된 집합을 잘 찾아내기도 하고, 특이발현 정도는 약하지만 같은 패턴을 보이는 다수의 유전자가 포함된 집합을 잘 찾는 등 결과에 차이가 많다. 그러므로 어떤 방법이 좋은지에 대한 답은 어떤 측면을 고려하는가에 따라 달라질 것이다. 본 연구에서는 유전자집합의 표현형에 따른 유의한 차이를 어떻게 재는가 보다는 유의성 정도를 나타내는 유의확률을 빨리 구하는 방법을 찾는 것이 목적이다.

논문의 2절에서는 표현형에 따라 유전자집합의 발현의 차이를 검색하는 보편적인 검정법을 제안하고 이 방법이 갖는 장단점을 논한다. 3절에서는 이 방법의 유의확률을 쉽게 구할 수 있는 대안들을 소개하고 실제 마이크로어레이 자료를 이용하여 4절에서 비교 분석한다.

2. 통계량 $\sum_{g\in G} t_g^2$와 순열검정

n개 표본의 마이크로어레이 자료가 있다. i번째 표본의 g번째 유전자 발현값을 X_{gi}, 그리고 표현형을 Y_i라 하자 ($g = 1, 2, \ldots, n_g$, $i = 1, 2, \ldots, n$). 표본의 표현형은 여러 형태가 있지만 여기서는 처리군 또는 비교군의 이진 표현형으로 국한한다.

p개의 유전자로 구성된 집합 G가 n_1개 표본의 처리군과 n_2 $(= n - n_1)$개 표본의 대조군 사이에 유의한 차이를 보이는지 알아보기 위한 척도는 우선 G에 속한 각 유전자들의 특이발현을 측정한 후 이들을 통합하는 것이다. 어떤 방법이 유의한 집합을 잘 찾아내는지 알아보기 위하여 Ackermann과 Strimmer (2009)는 유전자 수준의 특이발현을 재는 척도로 사용하는 통계량과 그 변환 방법, 그리고 G에 속하는 유전자들의 통계량을 통합하는 방법 등, 기존에 발표된 방법들을 조합한 261가지 방법을 비교하였다. 모의실험 결과, 유전자 수준의 특이발현을 측정하는 통계량의 선택은 검정력에 큰 영향을 끼치지 않지만 유전자별 통계량을 제곱하는 것이 가장 좋은 변환이라는 것을 보였다. 결론적으로 Ackermann과 Strimmer (2009)는 유전자집합의 유의성검정통계량으로 제일 간단한 각 유전자의 t-통계량의 제곱합 $\sum_{g\in G} t_g^2$을 사용할 것을 제안하였다. t-통계량도 여러 형태가 있으나 g번째 유전자의 처리군에서의 표본평균과 표본분산을 $\bar{X}_g^T$, $(S_g^T)^2$, 대조군에서는 $\bar{X}_g^C$, $(S_g^C)^2$라 할 때 두 군 사이의 특이발현을 재는 $t_g = (\bar{X}_g^T - \bar{X}_g^C)/\sqrt{(S_g^T)^2/n_1 + (S_g^C)^2/n_2}$의 사용이 일반적이다. 이 때 유전자 자료는 유전자별로 평균 0, 분산 1이 되도록 표준화하고 t-통계량 대신 유전자와 표현형 사이의 상관계수도 통계량으로 사용 가능하다.

유의성검정을 할 때는 주어진 정보를 토대로 검정통계량의 유의확률을 구하지만 검정통계량의 분포를 모른다면 관측된 자료로부터 가능한 많은 순열자료(permuted sample)를 생성하여 경험적인 분포를 만든 후 유의확률을 구할 수 있다. 앞에서 언급한 경쟁귀무가설에서는 집합에 속하는 유전자들을 임의로 바꿔주는 순열검정을 실시하고 자립귀무가설에서는 각 표본들의 표현형을 임의로 배정하는 순열검정을 한다. 그런데 유전자 치환의 경우에는 유전자들 사이의 상관관계, 특히 같은 집합에 속해 있는 유전자들 간의 관계가 무시된다는 맹점이 있고 본 연구에서는 관심 집합에 속한 유전자들의 발현 형태가 표현형과 연관성이 있는지에 초점을 맞추고 있기 때문에 표본의 표현형을 치환하는 것을 위주로 논하겠다.

전체 표본의 표현형을 랜덤치환 후 G에 속한 유전자들의 t-통계량의 제곱합, $(\sum t_g^2)^B$를 구한다. 이런 과정을 M번 반복하여 구한 값들을 실제 자료값 $\sum t_g^2$와 비교하여 아래와 같이 유의확률(p-value)을 계산한다.

$$p\text{-value} = \frac{\sum_{B_i=1}^{M} I\left(\left(\sum t_G^2\right)^{B_i} > \sum t_G^2\right) + 1}{M + 1}.$$

순열검정의 과정은 프로그램이 간단하며 어떤 복잡한 통계량에 대해서도 분포에 대한 가정없이 정확한 유의확률을 구할 수 있다는 장점이 있다. 그런데 치환을 M번 실시하여 얻을 수 있는 유의확률의 최소

값과 간격이 모두 $1/(M+1)$이다. 유의성검정을 실시하는 집합이 수백-수천 개인데 그들의 유의성 정도에 대한 변별력과 다중검정까지 고려한다면 M이 얼마나 커야할지 충분히 이해가 될 것이다. 또한 매번 생성된 자료들이 서로 다르기 위해서는 표본의 크기 n_1과 n_2도 따라서 커야할 것이다.

실제로 수만 개 유전자의 수백 개 유전자집합에 대한 유의성을 알아보기 위한 순열검정을 실행하는 컴퓨터 작업은 대용량의 메모리도 필요하지만 치환횟수에 비례한 작업실행 시간이 필요하다는 어려움이 있다. 본 연구에서는 검정통계량 $\sum t_g^2$의 유의확률을 쉽게 구할 수 있는 대안들을 찾아보았다.

3. 귀무가설 아래에서 통계량 $\sum_{g\in G} t_g^2$의 근사분포 설정

3.1. 중심극한정리를 이용한 방법

두 개의 서로 독립인 모집단에서 각 n_1, n_2개의 표본을 뽑아 t-통계량을 구하였을 때 표본크기가 충분히 크면 중심극한 정리에 의해 t-통계량의 분포는 정규분포로 수렴한다. 그리고 서로 독립이고 표준정규분포를 따르는 p개 통계량들의 제곱합은 자유도가 p인 카이제곱분포를 따른다는 것도 자명한 일이다. 그러므로 p개 유전자로 구성된 집합의 유의성 분석에서 두 군의 표본수가 크고 유전자들이 상호 독립이라면 유전자들의 t-통계량의 제곱합 $\sum t_g^2$도 기본적으로 자유도가 p인 카이제곱 분포에 근사하리라 여겨진다.

염색체 위치가 같거나 동일한 대사경로의 기능을 수행하는 유전자들 사이에는 상호관계가 존재하는게 당연하다. 그러므로 Tian 등 (2005)은 집합 G에 속하는 유전자들의 t-통계량의 합은 정규분포를 따르지 않고 유의확률 계산을 위해서는 순열검정을 실시할 수 밖에 없다고 하였다. 그러나 Irizarry 등 (2009)은 Mootha 등 (2004)과 Subramanian 등 (2005)에서 거론한 8개 자료를 분석하여 유전자의 t-통계량들, 그리고 유전자집합의 t-통계량의 합이 소수의 경우만 제외하고는 정규분포를 따른다는 것을 정규확률도를 통하여 보였고 실제 각 유전자집합에 속한 유전자들의 상관계수는 평균 0.1 수준임을 언급했다. 여기서 제외된 소수들이 바로 두 표현형의 차이를 나타내는 특이발현유전자이며 특이발현유전자집합인 것이다. 특히 Mootha 등 (2003)이 Kolmogorov-Smirnov 검정을 이용하여 찾아내 유명한 OXPHOS 유전자집합도 t-통계량의 합이 정규확률도에서 벗어난 이상점으로 금방 찾아낼 수 있음을 보였다. Irizarry 등 (2009)은 유전자집합의 유의성을 나타내는 검정통계량으로 아래와 같이 카이제곱 통계량을 표준화한 통계량을 제시하였다.

$$\frac{\sum(t_i-\bar{t}_G)^2-(p-1)}{\sqrt{2(p-1)}} \sim N(0,1), \quad \text{단, } \bar{t}_G=\sum_{i\in G}\frac{t_i}{p}.$$

3.2. 정규점수 변환을 이용한 방법

마이크로어레이 실험에서 오차를 발생하는 요인은 여러 가지가 있기 때문에 각 표본들을 대상으로 관찰값에 적절한 변환을 주어 비교 대상인 RNA 시료간 차이와는 무관한 체계적 변동원을 제거하는 표준화 과정(normalization process)이 반드시 필요하다. 특정 표본의 유전자 관찰값이나 각 유전자별 전체 평균값의 분포를 기준삼아 표본별 분포를 동일하게 하기도 하지만 표본별 발현값의 순위를 이용하여 표본의 분포를 통일하는 방법도 있다 (Bolstad 등, 2002; Zahn 등, 2006).

Zahn 등 (2006)은 표본별 유전자 발현값의 순위를 이용한 정규점수(normal score)를 관찰값 대신 사용하였다. 즉, i번째 표본에 속한 n_g개의 유전자 발현값들을 크기순으로 나열하여 g번째 관찰값 X_{gi}의 순위 r_{gi}를 구하고 이를 이용한 정규점수 $Z_{gi}=\Phi^{-1}((r_{gi}+0.375)/(n_g+0.25))$를 계산한다. 그러면 $\{Z_{gi}\}_{g=1}^{n_g}$은 표준정규분포를 따르고 이들로부터 처리군과 대조군 사이의 평균의 차이를 측정하는 이

표본 t-통계량 $\{t_g^{NS}\}_{g=1}^{n_g}$도 정규분포를 따르므로 유전자들이 서로 독립이라는 조건만 만족한다면 집합 G에 속하는 p개 유전자의 순위를 이용한 t-통계량의 제곱합 $\sum(t_g^{NS})^2$은 자유도가 p인 카이제곱분포를 따를 것이다.

순위를 이용한 정규점수변환 과정을 거치면 표본별 꼬리 부분 유전자들의 특별한 현상까지 제거된다는 지적도 있었지만 Bolstad 등 (2003)은 실증적으로 아무런 문제가 없음을 보였다. 또한 Tan 등 (2006)는 모의실험과 실제 자료분석을 통하여 순위를 이용한 분석이 표본크기가 작거나 발현값의 잡음이 많거나 변이가 큰 경우에 Significance Analysis of Microarrays (SAM) (Tusher 등, 2001)보다 훨씬 더 효율적임을 보였다.

3.3. 적률을 이용한 방법

표본의 표현형 $\{Y_i\}$과 발현값 $\{X_{gi}\}$가 아래 가정을 만족할 때 $\hat{\beta}_g = \sum X_{gi}Y_i/n$는 g번째 유전자와 표현형 사이의 표본공분산을 나타내며 유전자집합 G가 표현형과 독립인지 검정하는 통계량으로는 임의의 양수 w_g에 대해 $\hat{C} = \sum w_g\hat{\beta}_g^2$을 사용할 수 있다.

$$\sum_{i=1}^{n} Y_i = 0, \qquad \sum_{i=1}^{n} X_{gi} = 0, \qquad \sum_{i=1}^{n} X_{gi}^2 = n.$$

순열검정을 하여 관찰값 $\hat{C}$의 분포를 생성하는 대신 Larson과 Owen (2015)은 적률을 이용한 모수적 근사방법을 소개하였다. 치환된 표현형 $\{\tilde{Y}_i\}$로부터 계산한 검정통계량의 값을 $\tilde{C}$라 할 때 $\tilde{C}$의 근사분포로 카이제곱분포의 수정된 형태인 $\sigma^2\chi^2(\nu)$로 가정하고 σ^2와 ν의 값은 $\tilde{C}$와 $\sigma^2\chi^2(\nu)$의 $1, 2$차 적률을 이용하여 다음과 같이 구할 수 있다.

$$\sigma^2 = \frac{\mathrm{Var}\left(\tilde{C}\right)}{2E\left(\tilde{C}\right)}, \qquad \nu = \frac{2E\left(\tilde{C}\right)^2}{\mathrm{Var}\left(\tilde{C}\right)}.$$

Larson과 Owen (2015)은 $E(\tilde{Y}) = 0$, $E(\tilde{\beta}_g) = E(\sum X_{gi}\tilde{Y}_i/n) = 0$임을 이용하여 $\tilde{C}$의 기댓값과 분산을 구하였다.

$$\begin{aligned}
E\left(\tilde{C}\right) &= \sum_{g\in G} w_g E\left(\tilde{\beta}_g^2\right) = \frac{\mu_2}{n-1}\sum_{i=1}^{n} w_g\bar{X}_{gg}, \\
\mathrm{Var}\left(\tilde{C}\right) &= \sum_{g\in G}\sum_{h\in G} w_g w_h \mathrm{cov}\left(\tilde{\beta}_g^2, \tilde{\beta}_h^2\right) \\
&= \sum_{g\in G}\sum_{h\in G}\left[\frac{w_g w_h}{(n-1)(n-2)(n-3)}\begin{pmatrix}\mu_2^2\\ \mu_4\end{pmatrix}^T\begin{pmatrix} n^3-3n^2+3n & -3n^3+3n^2 \\ -n^2+n & n^3+n^2\end{pmatrix}\begin{pmatrix}\bar{X}^*_{gghh}/n^2 \\ \bar{X}_{gghh}/n^3\end{pmatrix}\right. \\
&\quad \left. - \frac{w_g w_h \mu_2^2}{(n-1)^2}\bar{X}_{gg}\bar{X}_{hh}\right]
\end{aligned}$$

단,

$$\begin{aligned}
\mu_2 = \frac{1}{n}\sum^{n} Y_i^2, \qquad \mu_4 = \frac{1}{n}\sum^{n} Y_i^4, \\
\bar{X}_{gh} = \frac{1}{n}\sum^{n} X_{gi}X_{hi}, \qquad \bar{X}_{ghrs} = \frac{1}{n}\sum^{n} X_{gi}X_{hi}X_{ri}X_{si}, \\
\bar{X}^*_{ghrs} = \bar{X}_{gh}\bar{X}_{rs} + \bar{X}_{gs}\bar{X}_{hr} + \bar{X}_{gr}\bar{X}_{hs}, \quad g,h,r,s \in G,
\end{aligned}$$

여기서 $w_g = (\mathrm{Var}(Y)^2 \cdot ((S_g^T)^2/n_1 + (S_g^C)^2/n_2))^{-1}$일 때 $\sum t_g^2 = \sum w_g \hat{\beta}_g^2$이므로 Larson과 Owen (2015)이 제시한 적률을 이용한 방법으로 귀무가설 아래에서 $\sum t_g^2$가 근사하는 수정된 형태의 카이제곱 분포를 찾을 수 있고 유의확률 계산도 가능한 것이다.

4. 근사방법에 따른 유의확률의 계산과 비교

p개의 유전자로 구성된 대사경로 G가 서로 다른 표현형 사이에 유의한 차이를 보이는지 검정하는 기본 통계량으로 Ackermann과 Strimmer (2009)는 $\sum t_g^2$가 가장 우수함을 보였는데 검정통계량의 분포가 확실하지 않을 때 유의확률 계산은 순열검정 M_1을 통하여 얻어질 것이다.

$$M_1 : \text{전체 표본의 표현형을 랜덤치환.}$$

순열검정의 단점들을 극복하기 위한 대안으로 3절에서 설명한 M_2, M_3, M_4, M_5의 모수적 방법으로 유의확률을 구할 수 있다.

$$\begin{aligned}
M_2 &: \sum t_g^2 \sim \chi^2(p), \\
M_3 &: \frac{\sum (t_i - \bar{t}_G)^2 - (p-1)}{\sqrt{2(p-1)}} \sim N(0,1), \\
M_4 &: \sum \left(t_g^{NS} \right)^2 \sim \chi^2(p), \\
M_5 &: \sum t_g^2 \sim \sigma^2 \chi^2(\nu).
\end{aligned}$$

이들을 비교하기 위한 유전자자료는 Bioconductor(www.bioconductor.org)에 data package로 실려 있고 Rahmatallah 등 (2014)에서 함께 분석한 'p53'과 'ALL'의 마이크로어레이 자료를 사용하였다. 또한 순열검정 M_1의 랜덤치환 횟수는 $M = 10^5$으로 하였다.

- p53 자료

  p53 단백질은 종양을 억제하는 기능을 하기 때문에 대부분의 암에서는 변이된 p53 유전자가 나타나고 분자생물학에서 이에 대한 연구가 활발하다. 유전자집합분석의 시발점이 되었던 Subramanian 등 (2005)와 Efron과 Tibshirani (2007) 등에서 다룬 p53 자료는 돌연변이군 33명과 정상군 17명의 10,010개 유전자가 담긴 마이크로어레이 자료이다.

- ALL 자료

  급성 림프구형 백혈병 관련 127개 표본, 12,625개 유전자의 마이크로어레이 자료로 구성되어 있고 Chiaretti 등 (2004, 2005)에서 이에 대한 자세한 분석을 하였다. 이중 B세포형 급성 림프구형 백혈병 환자이면서 BCR/ABL 돌연변이를 보이는 37명과 정상 72명의 자료를 분석하였다.

p53과 BCR/ABL의 돌연변이가 어떤 유전자집합에서 발현의 차이를 보이는지 알아보는데 사용한 905개 집합은 The Molecular Signatures Database (MSigDB)의 'C2:curated gene sets'에서 다운받았다. 유의확률을 계산하는데 사용된 방법들은 각 유전자들이 서로 독립이고 중심극한정리를 이용한 정규근사를 포함하고 있기 때문에 집합에 속한 유전자 수의 영향을 받을 수 있다. 이를 확인하기 위하여 905개 집합을 각 집합에 속한 유전자 수가 25개 이하이면 Group = 1, 26–50개이면 Group = 2, 51–100개이면 Group = 3, 그리고 100개를 초과하면 Group = 4의 네 그룹으로 분류하였다.

p53 자료와 ALL 자료를 대상으로 905개 집합의 유의확률을 구하였는데 M_1–M_4 방법으로는 10^{-10}보다 작은 값도 나왔지만, M_5의 최소값은 약 10^{-5}였다. Table 4.1은 R-프로그램을 이용하여 각 방법으

Table 4.1. Number of sets with p-value = 0 in p53 and ALL using M_1–M_5

		M_1	M_2	M_3	M_4	M_5
Number of sets	p53	0	1	8	1	0
with p-value = 0	ALL	0	168	357	183	0
Value substituted for p-value = 0		·	10^{-16}	10^{-16}	10^{-16}	·

Table 4.2. Correlations between the p-values of M_1–M_5 in p53 and ALL

	p53				ALL			
	M_2	M_3	M_4	M_5	M_2	M_3	M_4	M_5
M_1	0.9696	0.7408	0.9642	0.9855	0.9532	0.6739	0.8972	0.9976
M_2	1	0.6902	0.9911	0.9726	1	0.5903	0.9143	0.9511
M_3		1	0.6863	0.7174		1	0.5541	0.6852
M_4			1	0.9681			1	0.8809

로 구한 출력물의 유의확률이 0인 집합의 수를 나타낸 표인데 p53 자료에서는 M_2와 M_4를 사용한 경우 1개, M_3로는 8개가 발견된 것에 비해 ALL 자료에서 M_2–M_4의 방법을 적용하였을 때 유의확률이 0인 집합도 많았고 전체적으로 50% 이상이 10^{-8}이하의 유의확률을 가졌다. 이것은 ALL 자료에서 어느 한 집합에라도 속한 유전자들 중 20%가 2 이상의 t-통계량 절대값을 가졌기 때문일 것이다.

Table 4.2는 p53 자료와 ALL 자료에서 M_1–M_5의 방법으로 구한 유의확률들의 상관계수 표이다. 전체적으로 보았을 때 카이제곱 통계량을 표준화한 M_3는 다른 방법들과 차이를 보임을 알 수 있다. 순열검정을 실시한 M_1의 값을 기준으로 비교하였을 때 M_3는 M_1을 대신하기에 적당하지 않고 유전자들이 서로 독립이라는 가정 아래 진행된 M_2, M_4보다는 적률을 이용하여 수정된 카이제곱분포를 도입한 M_5가 더 좋은 대안이라고 판단된다. M_2와 M_4는 분석자료로 유전자 발현값을 직접 이용하는지, 아니면 발현값의 순위에 따라 정규점수로 변환한 값을 이용하는지 차이가 있을 뿐, 모두 t-통계량의 제곱합을 사용하고 각 유전자가 독립이라는 가정 아래 카이제곱 분포로부터 각 통계량의 유의확률을 구한 결과이다. 논문에 산포도를 싣지는 않았지만 M_2와 M_4의 유의확률은 크기에 약간의 차이가 있을 뿐, 유의확률이 0에 가까울 때 각 집합의 유의한 순위가 완전 일치함을 볼 수 있었다. 다만 ALL 자료는 특이발현 유전자가 많다는 특이점으로 통계량 값의 변동이 커서 p53자료에 비해 두 방법의 상관계수가 약간 작은 현상을 보였다.

M_1–M_5 방법을 이용하여 구한 유의확률들을 비교하는데 우리 관심의 대상인 0에 가까운 부분을 자세히 보기 위하여 '$-\log_{10}$' 변환을 실시하였고 M_2–M_4의 유의확률이 0인 경우는 로그변환을 위하여 10^{-16}의 값으로 대체하였다. Figure 4.1은 p53의 자료, Figure 4.2는 ALL 자료를 대상으로 905개 집합의 유의확률을 비교한 그림이다. 각 그림의 4개 그림들은 수평축에는 M_1의 방법으로 얻은 $-\log_{10}$(유의확률), 수직축에는 M_2–M_5 각각의 방법으로 얻은 $-\log_{10}$(유의확률)의 산포도로 각 집합의 유전자 수를 나타내는 변수 Group에 따라 각 점의 색과 모양을 다르게 표현하였다.

모든 산포도의 기울기는 1보다 큰데 M_5보다는 M_2–M_4에서, 그리고 Group 변수가 큰 값을 가질수록 이 현상이 두드러짐을 볼 수 있다. 같은 집합에 속하여 동일한 기능을 수행하는 유전자들은 서로 연관성이 높기 때문에 독립성을 가정한 M_2–M_4 결과에서는 집합에 속한 유전자수가 많을수록 독립성 가정에 위배되는 정도가 심해지므로 M_1과 일치도가 떨어짐을 볼 수 있다. 다른 방법들과는 달리 M_3는 유의확률이 0을 벗어난 경우에도 M_1과 일치도가 떨어지는데 이것은 M_3가 독립성 결여 외에도 표준화를 이용하여 무리한 정규근사를 시도한 때문이라 생각된다. Table 4.2의 상관계수 비교로부터 예견되었듯이 M_1을 대신할 방법으로는 M_5가 제일 바람직함을 확인할 수 있다.

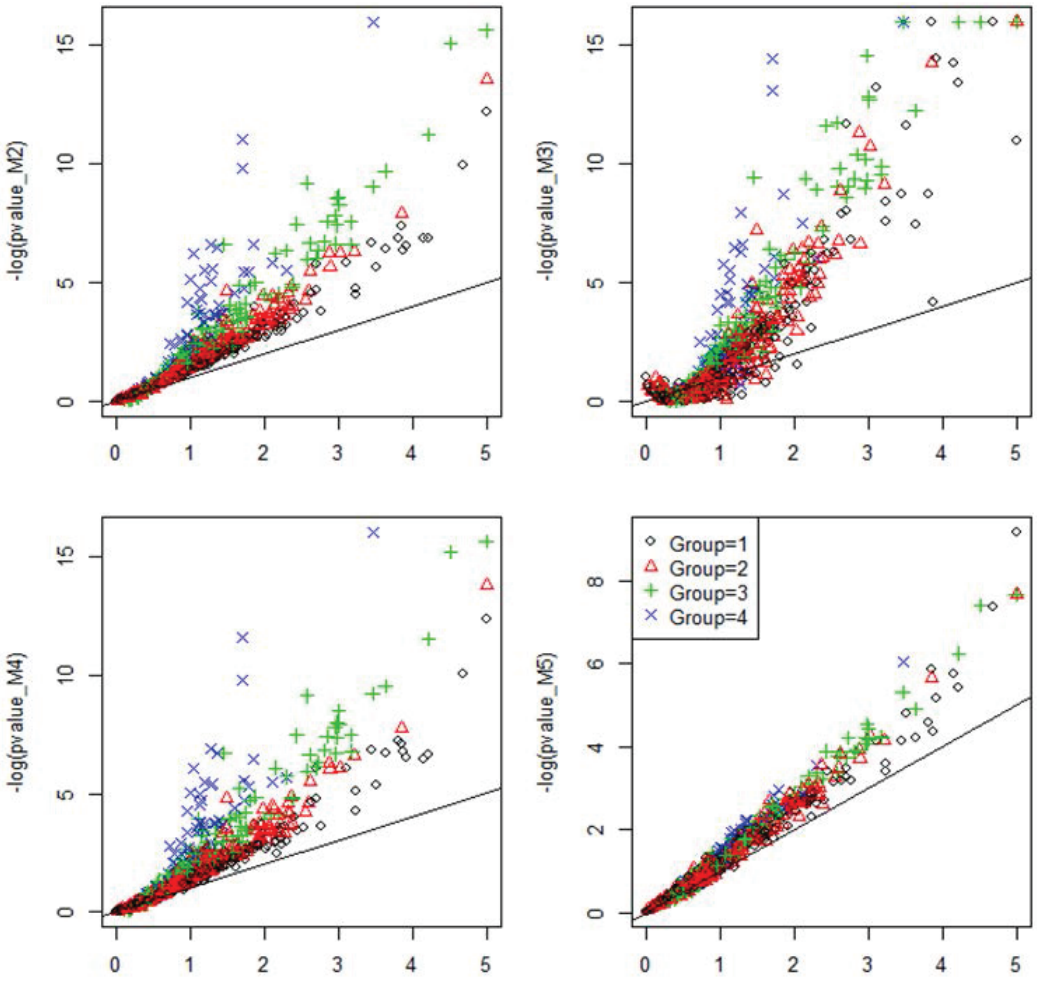


Figure 4.1. Four scatter plots of $-\log(p\text{-value})$ using M_1 vs. $-\log(p\text{-value})$ using others(M_2–M_5) for 905 gene sets of p53 data.

905개 집합의 검정이 동시에 실시되었기 때문에 다중검정을 위한 제1종 오류 보정은 필수적이다. Bonferroni (1936)의 family wise error rate 같은 너무 엄격한 보정 대신 유전자 분석에서 많이 이용되는 Benjamini와 Hochberg (1995)와 Yekutieli와 Benjamini (1999)의 false discovery rate 보정인 q-value를 계산하였다. Figure 4.3은 p53과 ALL 자료에서 얻은 905개 q-vlaue를 비교한 산포도인데 수평축은 M_1의 방법으로 얻은 $-\log_{10}(q\text{-value})$, 수직축은 M_5 방법으로 얻은 $-\log_{10}(q\text{-value})$를 나타낸다. 이 산포도를 보면 표본치환의 횟수에 의존하는 M_1의 문제점이 확연히 들어난다. M_1의 유의확률 최소단위는 10^{-5}였기 때문에 q-value의 연속성 결여가 눈에 뜨이고 특히 돌연변이 군과 정상군 사이에 유의한 차이를 보이는 집합이 많았던 ALL 자료의 경우에는 M_1의 방법으로 얻은 q-value로는 변별력이 없음을 볼 수 있다.

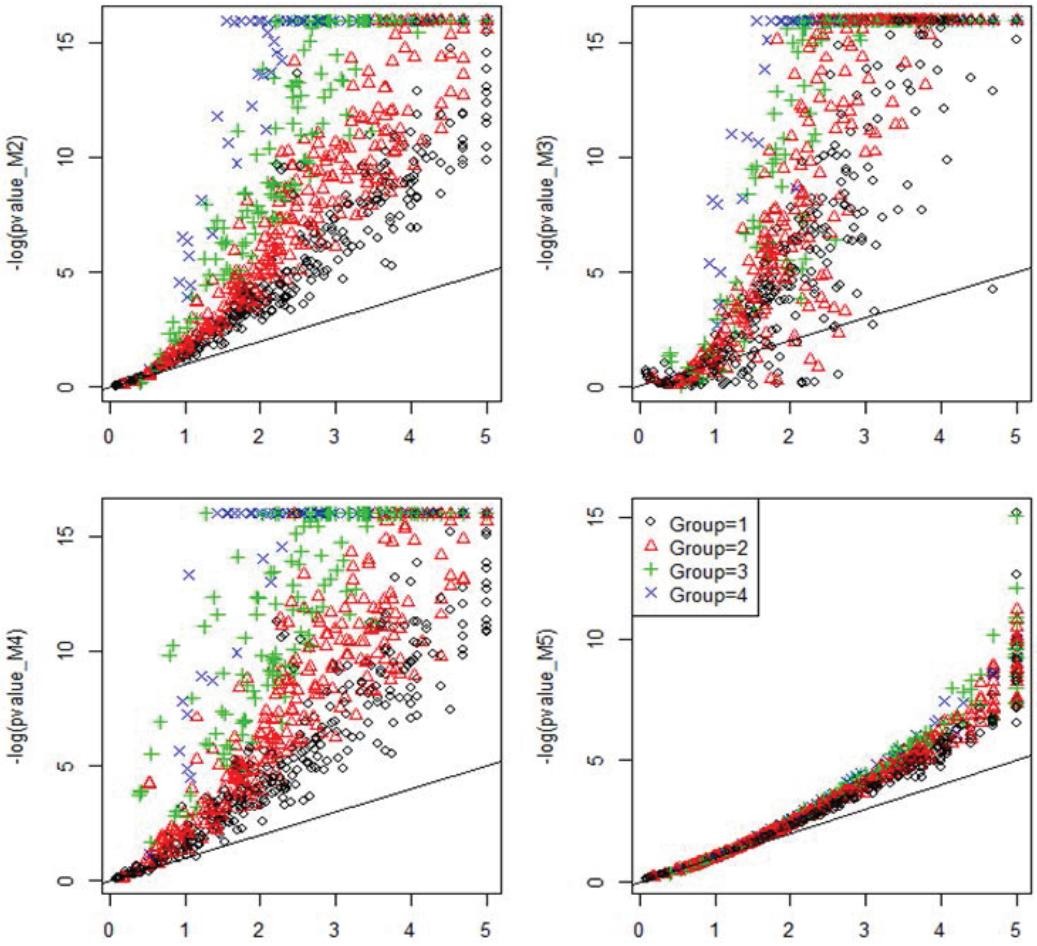
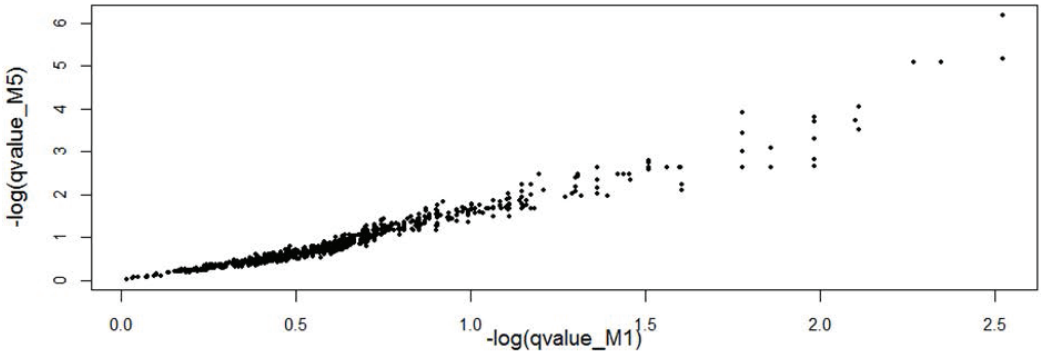


Figure 4.2. Four scatter plots of $-\log(p\text{-value})$ using M_1 vs. $-\log(p\text{-value})$ using others(M_2–M_5) for 905 gene sets of ALL.

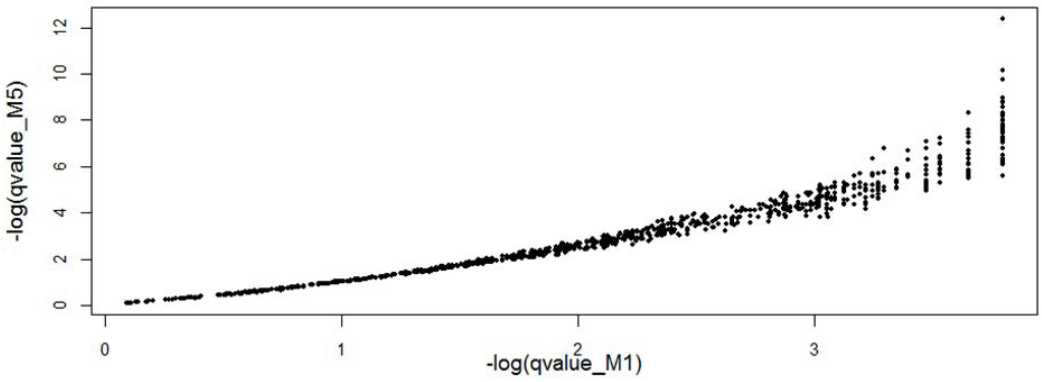
5. 결론

마이크로어레이 자료의 유전자집합분석에서 기존에 발표된 통계량들을 모의실험을 통하여 비교한 결과, Ackermann과 Strimmer (2009)는 집합에 속한 유전자들의 t-통계량의 제곱합 $\sum t_g^2$이 간단하지만 이진 표현형 사이에 유의한 차이를 보이는 집합을 제일 잘 찾아내는 검정통계량이라 밝혔다. 검정 대상 유전자집합의 유의확률 계산을 위해서는 순열검정을 실시하는 것이 기본이고, 비현실적이지만 동일한 기능을 수행하는 유전자들이 서로 독립이라고 가정하면 모수적 방법으로 유의확률을 쉽게 구할 수 있다. Mootha 등 (2003)의 GSEA는 유전자들 사이의 연관성을 이용해 집합의 유의성을 검정한 방법이고 Kim과 Volsky (2005)의 PAGE는 유전자의 독립성과 중심극한정리를 이용한 모수적 검정 방법인데 이 두 가지 대표적인 방법에 대한 득과 실은 명확한 차이가 있다.

$\sum t_g^2$을 사용한 유전자집합의 유의성검정에서 M_1는 유전자들 사이의 연관성을 그대로 반영하여 순열검정으로 유의확률을 구한 방법이고 M_2–M_4는 모든 유전자들이 서로 독립이라는 가정아래 카이제곱분포



(a) p53



(b) ALL

Figure 4.3. Scatter plots of $-\log(q\text{-value})$ using M_1 vs. $-\log(q\text{-value})$ using M_5 in p53 and ALL.

를 이용한 검정이다. 이에 비해 M_5는 $\sum t_g^2$가 수정된 형태의 카이제곱분포를 따른다고 가정하고 유의확률을 계산하였다. 이 과정에서 M_5는 유전자들이 서로 독립이라고 가정하지도 않았고 계산 시간이 많이 걸리는 표본치환의 과정을 간단한 적률 계산으로 대신하였다.

컴퓨터 사양에 따라 계산 속도가 다르지만 10,000개 유전자들을 대상으로 905개 집합의 유의확률을 구할 때 대략 시간 당 10,000번 정도 치환한 순열검정이 가능하였고 카이제곱분포를 이용한 유의확률 계산은 2–3초 정도, 그리고 적률을 이용한 계산은 1–2분 정도 걸렸다. 검정통계량의 분포를 알지 못할 때 유의확률을 구할 수 있는 정확한 방법은 분명 순열검정을 실시하는 것이고 제대로 된 순열검정을 위해서는 500,000개 정도의 치환표본을 생성하라고 하지만 컴퓨터의 용량과 계산에 걸리는 시간을 생각하면 쉽게 얻을 수 있는 결과가 아니다. 모수적 검정인 M_2–M_4는 컴퓨터 프로그램을 실행하자마자 결과를 얻을 수 있는 반면, 수정된 카이제곱분포를 이용한 M_5는 계산을 순간적으로 해내지는 못하지만 컴퓨터 앞에 그대로 앉아 결과를 기다릴 수 있는 시간이다.

실제 자료분석을 통하여 905개 집합의 유의확률들을 비교한 결과, M_5는 M_1에 비해 좀 작은 유의확률을 제시하는 경향이 있으나 단 시간에 계산이 가능하며 유의확률이 아주 작은 경우 M_1과 순위가 거의 일치한 근사값을 제시하였다. 또한 M_1이 변별력을 보일 수 없는 경우에도 M_5는 훨씬 작은 유의확률

값을 계산해냄으로서 다중검정을 실시한 이후에 유의한 유전자집합들을 추가로 발견할 수 있었다. 적률 계산을 이용한 M_5가 계산 시간도 훨씬 짧고 아주 작고 세밀한 유의확률도 계산 가능한 장점을 지닌, 순열검정의 훌륭한 대안임을 확인하였다.

References

Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis, *BMC Bioinformatics*, **10**, 47.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, **19**, 185–193.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilitá, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival, *Blood*, **103**, 2771–2778.

Chiaretti S., Li, X., Gentleman, R., Vitale, A., Wang, K. S., Mandelli, F., Foa, R., and Ritz, J. (2005). Gene expression profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation, *Clinical Cancer Research*, **11**, 7209–7219.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes, *The Annals of Applied Statistics*, **1**, 107–129.

Irizarry R. A., Wang, C., Zhou, Y., and Speed, T. P. (2009). Gene set enrichment analysis made simple, *Statistical Methods in Medical Research*, **18**, 565–575.

Kim, S.-Y. and Volsky, D. J. (2005). PAGE: Parametric analysis of gene set enrichment, *BMC Bioinformatics*, **6**, 144.

Larson, J. L. and Owen, A. B. (2015). Moment based gene set tests, *BMC Bioinformatics*, **16**, 132.

Mooney, M. A. and Wilmot, B. (2015). Gene set analysis: a step-by-step guide, *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics*, **168**, 517–527.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., *et al.* (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature Genetics*, **34**, 267–273.

Rahmatallah, Y., Emmert-Streib, F., and Glazko, G. (2014). Gene sets net correlations analysis (GSNCA): a multivariate differential coexpression test for gene sets, *Bioinformatics*, **30**, 360–368.

Subramanian, A., Tamayo, P., Mootha, V. K., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. In *Proceedings of the National Academy of Sciences*, **102**, 15545–15550.

Tan, Y. D., Fornage, M., and Fu, Y. X. (2006). Ranking analysis of microarray data: a powerful method for identifying differentially expressed genes, *Genomics*, **88**, 846–554.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. In *Proceedings of the National Academy of Sciences*, **102**, 13544–13549.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. In *Proceedings of the National Academy of Sciences*, **98**, 5116–5121.

Yekutieli, D. and Benjamini, Y. (1999). Resampling based false discovery rate controlling multiple test procedure for correlated test statistics, *Journal of Statistical Planning and Inference*, **82**, 171–196.

Zahn, J., Sonu, R., Vogel, H., *et al.* (2006). Transcriptional profiling of aging in human muscle reveals a common aging signature, *PLoS Genetics*, **2**, e115.

유전자집합분석에서 순열검정의 대안

이선호[a,1]

[a]세종대학교 수학통계학부

(2018년 1월 23일 접수, 2018년 2월 21일 수정, 2018년 2월 26일 채택)

요 약

마이크로어레이 자료의 유전자집합분석은 개별유전자분석에 비해 검정력도 높일 수 있고 결과 해석이 쉬워서 이에 대한 연구가 활발히 진행되어 왔다. 표현형에 따라 유의한 차이를 보이는 유전자집합의 검색은 검정통계량들이 유도된 배경에 따라 결과에 차이를 보이지만 대체적으로 t-통계량의 제곱합을 이용한 순열검정이 제일 무난한 방법으로 여겨진다. 그러나 유전자집합분석에서 다중검정은 필수이고 많은 집합들의 유의성에 변별력을 주기 위해서는 순열검정에서 생성하는 치환표본의 수가 많이 필요하고 시간이 오래 걸린다는 문제점이 있다. 순열검정을 대신할 모수적 방법들을 검토한 결과, 적률을 이용한 근사가 각 집합의 유의확률 계산시간도 훨씬 단축하며 순열검정에서 구한 유의확률과 크기와 순위가 거의 일치함을 확인하였다.

주요용어: 다중검정, 순열검정, 유전자집합분석, 적률

[1](05006) 서울시 광진구 능동로 209, 세종대학교 수학통계학부. E-mail: leesh@sejong.ac.kr