응용통계연구 (2010)
23(6), 1067–1080
DOI: 10.5351/KJAS.2010.23.6.1067

고차원 자료의 재현성과 표본 수

서원석[1] · 최지아[2] · 정형철[3] · 조형준[4]

[1]고려대학교 통계학과, [2]고려대학교 통계학과, [3]수원대학교 통계정보학과, [4]고려대학교 통계학과

(2010년 8월 접수, 2010년 8월 채택)

요 약

임상시험을 위한 표본 수 산정방법에 대해 지금까지 많은 방법이 개발되었고 현재 국내외 임상시험 기관에서 이 방법들을 토대로 표본 수를 산정하고 있다. 하지만 마이크로어레이칩을 이용한 실험에 필요한 표본 수 산정에 대한 연구는 아직 미비하여 제대로 이용되지 않고 있다. 본 연구의 목적은 마이크로어레이 실험에 필요한 표본 수를 산정하는 데 있어 실제 마이크로어레이 자료의 재현성에 대한 정보를 이용하여 그 지침을 제공하는데 있다. 재현성 비교에서는 5가지 검정방법 즉, Fold change, Two-sample t-test, Wilcoxon rank-sum test, SAM, LPE 방법 별로 재현성을 측정하였다. 발현 값의 표준화 방법에 있어서는 MAS5, RMA 두 가지로 세분화 하였으며 반복수에 따라 상위 20개 또는 100개 유전자에 대한 일치성도 측정하였다. 또한, 표본수를 산정하는데 있어 기존에 제시한 방법에 현실적인 정보를 이용하여 좀 더 세분화하여 실험에 필요한 표본수를 산정해 보았다.

주요용어: 마이크로어레이, 재현성, 표본수, 효과의 크기.

1. 서론

현대사회가 급속도로 발전하면서 우리는 지금 이 시간에도 축적되고 있는 거대한 데이터를 축적하고 분석, 예측할 수 있는 기술을 가지게 되었다. 그러나 임상시험(clinical trial)이나 마이크로어레이(microarray)분석에 필요한 자료를 얻기 위해서는 윤리적, 도덕적인 부분이 존재하게 되며 실험의 고비용 문제 때문에 상대적으로 적은 데이터를 이용하게 된다. 더군다나 마이크로어레이 자료는 적게는 2개에서 많게는 50개 안팎의 데이터를 이용하기 때문에 기존에 비해 진보된 개념을 이용하여 통계적 방법을 적용해야 한다. 마이크로어레이는 고차원 자료(high-dimensional data)이며 어떤 생물체가 가지고 있는 수천, 수만 개의 유전자(gene)들의 발현값(expression value)을 한꺼번에 측정한다는 특징을 가지고 있기 때문에 기존과는 다른 방식으로의 접근이 요구된다. 마이크로어레이 자료와 같은 고차원 데이터를 생산하는 기술은 불과 수년 전에 개발되었기 때문에 마이크로어레이 실험에 있어 필요한 표본 수 계산에 대한 연구는 국제적으로 많지 않으며, 국내에서는 전무한 실정이다.

본 논문에서는 유의한 유전자 탐색 방법 중 5가지 검정방법(Fold change, Two-sample t-test, Wilcoxon rank-sum test, SAM, LPE)을 이용하여 6가지 실제 마이크로어레이 실험 자료에 대한 재현성(reproducibility)을 비교하고자 한다. 이때, 마이크로어레이 실험 자료에서는 두 가지 조건하에서 상이하게 발현되는 유전자(differentially expressed gene)를 찾기 위해 다중가설검정(multiple hypothesis testing)을 수행하게 되는데 수 만개나 되는 유전자를 동시검정 함으로 인해 심각한 다중가설검정 문제가 발

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2009-0087564).
[4]교신저자: (136-701) 서울시 성동구 안암동 5가, 고려대학교 통계학과, 교수. E-mail: hj4cho@korea.ac.kr

생한다. 다중가설검정에 있어서 제 1종 오류를 조정하는 기법에 대한 연구는 Benjamini와 Hochberg (1995), Benjamini와 Yekutieli (2001) 등이 있으며 Dudoit 등 (2003)은 마이크로어레이 자료의 분석에 적용되는 다양한 다중가설검정 기법들에 대한 비교논문을 발표하였다. Jung (2005), Shao와 Tseng (2007)는 다중가설검정에 있어서 FDR(False Discovery Rate)을 고려한 표본 수 산정 방법에 대한 연구가 있었다. 본 논문에서는 기존 연구에 비해 다양한 효과크기를 고려한 표본 수를 제시 하며 실제 데이터를 이용하여 재현성을 고려한 표본 수 계산방법도 제안하고자 한다.

본 논문에서 전개될 내용을 다음과 같이 구성하였다. 2장에서는 마이크로어레이 자료 분석에서 유의한 유전자를 탐색하는 기법과 다중가설검정 문제에서의 오류 비율의 조정(error rate control)에 대한 개념을 소개한다. 3장에서는 실제 자료들에 대한 재현성의 결과를 정리하며, 4장에서는 실제 자료를 이용하여 다양한 조건하에서의 표본 수 산정 결과를 제시한다. 끝으로 5장에서는 결론 및 앞으로의 연구방향에 관하여 논의한다.

2. 유전자 탐색과 다중검정 문제

2.1. 유의한 유전자 탐색 방법

2.1.1. Fold change 생물학 에서는 유의한 유전자를 탐색하는 방법으로서 계산과 해석 과정이 다른 통계적 방법에 비해 간편한 fold change를 많이 사용한다. Fold change는 두 그룹 간 발현값의 비(ratio)로서 정의되며 이 비가 2보다 크거나 0.5보다 작은 유전자는 유의한 것으로 판단한다(2-fold change). 그러나 이 방법은 유전자 간의 변동을 무시한 후 유의한 유전자를 탐색하기 때문에 잘못된 판단을 내릴 수 있는 가능성을 다수 내포한 방법이다.

2.1.2. Two-sample t-test 대표적인 모수적(parametric)방법으로서 가장 널리 쓰이는 기본적인 통계적 검정방법이다. Fold change방법과는 달리 변동(variation)을 고려하여 두 처리집단 간 유의한 유전자를 탐색하게 된다. 실질적으로 두 처리집단 간 등분산을 가정하지 않기 때문에 수정된 t-test, 즉 Welch's t-test를 사용하는 것이 일반적이다. 그러나 마이크로어레이 자료의 고비용 문제로 인해 2∼3개의 자료만을 이용하게 되는 경우가 발생하며 이 때 일반적인 t-test방법을 적용하게 되면 낮은 통계적 검정력을 가지게 된다. Jain 등 (2003)은 적은 자료를 이용하여 검정을 실시하여도 검정력(power)을 유지할 수 있는 방법(LPE)을 제안하였다.

2.1.3. Wilcoxon rank-sum test 이 방법은 분포의 가정을 필요로 하지 않는 비모수적(nonparametric)방법으로서 발현값의 순위를 이용하여 유의한 유전자를 탐색하게 된다. 이 검정통계량의 값이 클수록 두 처리집단 간 순위합의 차이가 크다는 것을 의미하며 그것은 곧 유의한 유전자임을 의미한다.

2.1.4. SAM(Significance Analysis of Microarrays) 귀무가설의 기각여부에 결정적인 영향을 미치는 검정통계량은 마이크로어레이 자료구조상 분모의 변화가 큰 경우를 볼 수 있다. 마이크로어레이 실험에서는 수 만개의 유전자에 대해 검정을 하기 때문에 분모가 상대적으로 작은 경우가 확률적으로 발생하므로 잘못된 결론을 내리는 경우가 많다. Tusher 등 (2001)은 분모의 안정화를 위해 수정인자(fudge factor)를 계산하여 대입하는 방법을 제안하였다. $d(i)$ 통계량(relative difference)을 구하는 식은 아래와 같다.

$$d(i) = \frac{\overline{x_I}(i) - \overline{x_U}(i)}{s(i) + s_0},$$

표 **2.1.** 가설검정 결과표

참가설	판단한 가설	
	기각하지 않음	기각함
H_0	True Positive (TP)	False Positive (FP)
H_1	False Negative (FN)	True Negative (TN)

$$s(i) = \sqrt{\left(\frac{1/n_1 + 1/n_2}{n_1 + n_2 - 2}\right)\left(\sum_m [x_m(i) - \overline{x_I}(i)]^2 + \sum_n [x_n(i) - \overline{x_U}(i)]^2\right)},$$

여기서 $\overline{x_I}(i)$, $\overline{x_U}(i)$은 그룹 I와 U에서의 유전자 i에 대한 평균 발현값을 나타내며 n_1, n_2는 각각 그룹 I와 U에서의 관측치 수를 나타낸다.

2.1.5. LPE(Local Pooled Error) 마이크로어레이 실험은 고비용 문제로 인해 아주 적은 표본으로 유의한 유전자를 식별해야 하는 경우가 많은데 이는 검정력이 낮아지는 문제점을 발생한다. Jain 등 (2003)은 적은 표본을 이용하여 검정을 시행하여도 통계적 검정력을 높이 유지할 수 있는 LPE 방법을 제시하였다.

$$z = \frac{\mathrm{Med}_1 - \mathrm{Med}_2}{\sigma_{pooled}}, \qquad \sigma^2_{pooled} = \frac{\pi}{2}\left[\frac{\sigma_1^2 \mathrm{Med}_1}{n_1} + \frac{\sigma_2^2 \mathrm{Med}_2}{n_2}\right],$$

여기서 $\mathrm{Med}_1, \mathrm{Med}_2$는 각각 첫 번째 표본과 두 번째 표본의 발현 강도 중위수(median intensity)를 나타낸다. LPE는 비슷한 발현강도를 갖는 유전자는 유사한 분산을 가지는 현상을 이용하여 근사기저분산을 추정, 분산 추정의 정밀도를 향상시키는 방법이다.

2.2. 다중검정 문제와 오류율의 조정

마이크로어레이와 자료는 가지고 있는 유전자의 개수만큼 검정을 함으로서 굉장히 많은 검정횟수와 더불어 제 1종 오류가 증가한다. 마이크로어레이에서의 제 1종 오류란, "두 집단간 발현값이 차이가 없다는 귀무가설이 맞는데도 귀무가설을 기각하여 발현값의 차이가 존재한다"라고 판단할 오류이다. 유전자의 개수가 늘어날수록 제 1종 오류가 증가하기 때문에 한 개의 유전자를 검정할 때 정한 유의수준을 그대로 사용하게 되면 잘못 판단한 유전자가 다수 포함되게 된다. 이러한 다중검정문제를 해결하기 위해 FWER(Family-Wise Error Rate)과 FDR을 이용하게 되는데 본 논문에서는 FDR의 개념을 적용하기로 한다.

2.2.1. FWER(Family-Wise Error Rate) FWER은 단일 검정에서의 유의수준 개념을 다중가설검정인 경우로 확장했을 때 다중비교에서의 유의수준 적용의 문제로 확장한 개념이다. FWER은 하나 이상의 검정을 동시에 수행할 때 적어도 하나가 제 1종 오류를 범할 확률로 정의된다. 표 2.1은 가설검정 결과표이다.

결국 FWER은 적어도 하나가 제 1종 오류를 범할 확률을 유의수준 보다 작도록 즉, $P(\mathrm{FP} \geq 1) \leq \alpha$가 되도록 조정하는 것이다.

2.2.2. FDR(False Discovery Rate) FWER은 여러 개의 검정을 동시에 할 때 제 1종 오류의 확률을 사전에 정의 된 유의수준 이하로 제어하는 방법이기 때문에 보수적(conservative)성격을 띤

다. FWER은 제 1종 오류가 발생할 가능성을 최소화하였기 때문에 검정력은 그만큼 낮아진다. Benjamini와 Hochberg (1995)는 제 1종 오류의 확률을 제어하고 FWER보다 검정력을 개선할 수 있는 FDR이라는 새로운 개념을 제시하였다. 이 개념은 유의한 검정결과 중 제 1종 오류를 범한 결과가 포함되는 비율을 제어하는 개념으로서 아래와 같이 정의된다.

$$\mathrm{FDR} = E\left(\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}\right),$$

여기서 FP는 0보다 크며, 만약 FP가 0이라면 FDR은 0이 된다. FDR은 FWER보다 덜 보수적인 방법이지만 유의한 검정의 개수가 많을수록 잘못 판단한 검정의 개수가 많아진다. 예를 들어 1,000개의 검정이 유의한 결과가 도출된다면 유의수준을 0.01로 하여도 10개의 검정이 평균적으로 잘못 판단한 것이기 때문이다. 따라서 위의 개념을 이용한 다양한 방법으로 FDR을 추정하게 된다.

1) Benjamini와 Hochberg의 FDR 추정 방법

예를 들어 m개의 유전자를 독립적으로 검정한다고 하고 검정결과 m개의 순서정렬된 p-value는 $P_{(r_1)},\ldots,P_{(r_m)}$라 하자. 이 때, 주어진 유의수준 α하에서 아래의 식을 만족하는 가장 큰 j까지의 p-value, 즉 $P_{(r_1)},\ldots,P_{(r_j)}$를 기각하는 방법이다.

$$P_{(r_j)} \le \frac{j}{m}\alpha, \quad j = 1,\ldots,m.$$

2) SAM의 FDR 추정 방법

SAM 방법을 통하여 도출된 $d(i)$를 이용하여 유의한 유전자를 탐색할 때 Tusher 등 (2001)은 순열(permutation)자료를 생성하여 경험적 분포(empirical distribution)을 추정한 후 FDR을 조정하는 방법을 제시하였다. 먼저 $d(i)$통계량의 기각역을 결정하면 FP + TN개의 기각된 유전자수가 도출되고, 각 B개의 순열자료 별 기각 유전자수 V_p, $p = 1,\ldots,B$가 계산된다. 그리고 각 순열자료 별로 $V_p/(\mathrm{FP}+\mathrm{TN})$를 계산하여 FDR값을 추정하게 되고 유의한 유전자를 탐색하게 된다.

3) LPE의 FDR 추정 방법

본 논문에서는 LPE의 FDR을 추정하기 위해 RIR(Rank-Invariant Resampling)을 적용하였다. 이 방법은 적은 표본으로도 신뢰성 있는 FDR을 추정하는데 있어 효과적이다 (Jain 등, 2005).

3. 실제 자료를 이용한 재현성 측정

3.1. 재현성 측정의 방법

우선 본 논문에서는 6개의 실제 마이크로어레이 실험 자료들 각각에 대해 두 조건하에서 상이하게 발현되는 유전자를 찾아내는 검정방법들인 Fold Change, Two-sample t-test, SAM, Wilcoxon rank-sum test, LPE를 이용하여 재현성을 측정하고 비교해 보았다. 우선 마이크로어레이 실험 분석 시 이용된 총 표본($n = n_1 + n_2$)에 대해 두 처리집단 간 5가지 방법으로 가설검정을 하여 검정통계량을 산출하였다. 즉, 마이크로어레이가 m개의 유전자를 내포하고 있으므로 m개의 검정통계량이 산출된다. 또한 각 처리집단 별로 $n'(< n_1, n_2)$개를 무작위로 비복원 추출하여 위와 동일하게 m개의 검정통계량을 산출한 후 Pearson의 상관계수와 Spearman의 상관계수를 모든 가능한 조합의 수만큼 평균과 표준오차를 계산한다. 만약 위 5가지의 방법들 중 재현성이 비교적 뛰어난 방법이 존재한다면 마이크로어레이 자료 분석

표 **3.1.** 각 데이터 별 상관관계

	MAS 5			RMA		
	Within	Between	Diff.	Within	Between	Diff.
T-cell	0.9464(0.008)	0.8841(0.006)	0.0623	0.9961(0.001)	0.9476(0.005)	0.0485
Prostate	0.8242(0.003)	0.8183(0.003)	0.0059	0.8716(0.005)	0.8643(0.005)	0.0073
HIV	0.9058(0.004)	0.9007(0.003)	0.0051	0.9366(0.004)	0.9321(0.003)	0.0045
Clozapine	0.9162(0.002)	0.9158(0.002)	0.0004	0.9461(0.002)	0.9453(0.002)	0.0008
Malaria	0.8354(0.002)	0.8238(0.002)	0.0116	0.9121(0.003)	0.8983(0.003)	0.0138
Head	0.8423(0.002)	0.8241(0.002)	0.0182	0.8892(0.002)	0.8753(0.002)	0.0139

에 있어서 이 방법은 표본 수 증가에 민감하지 않은, 변동이 적은 방법이라 할 수 있다.

또한 위의 5가지 방법들에 의해 산출된 통계량들의 상위 20개를 비교한다. 여기서 비교의 방법은 마이크로어레이 실험의 분석에 이용된 총 표본수에 대한 각 방법들의 상위 20개와 가능한 표본조합들을 고려한 상위 20개간에 일치하는 개수에 대한 평균과 표준오차를 제시하는 것이다. 위와 같은 방법으로 상위 100개의 유전자에 대해서도 평균과 표준오차를 제시하였다.

마이크로어레이 실험은 하나의 슬라이드를 통해 수 만개의 유전자를 한꺼번에 실험하기 때문에 실험과정에서 발생하는 여러 오차로 인해 실험의 정도가 동일하지 않을 수 있다. 따라서 발현 된 값들을 표준화 시킬 필요가 있는데 발현값에 대한 표준화 방법은 Affymetrix의 MAS5와 Irizarry 등 (2003)의 RMA(Robust Multichip Average)가 있으며 본 논문에서도 이 두 가지 방법들을 이용하였다.

3.2. 실제 마이크로어레이 데이터의 재현성

본 논문에서는 각기 다른 6개의 실험을 통해 재현성을 측정한 결과를 제시하였다. 마이크로어레이는 Affymetrix사의 GeneChip®을 이용하였다. 각 실험에서 사용된 표본 수, 유전자의 개수, 칩에 대한 정보는 홈페이지(http://www.korea.ac.kr/~stat2242/pub/)에 제시되어 있다.

표 3.1은 각 자료 별로 그룹-내(within), 그룹-간(between) 상관관계의 평균과 표준오차 그리고 그룹-내와 그룹-간의 상관관계의 차이(Diff)를 보여주고 있다. 상관관계가 모두 0.8이상이고 그룹-내 상관관계가 그룹-간 상관관계 보다 높으므로 대체로 실험이 잘 되었다고 할 수 있다. 또한 RMA방법이 MAS5보다 상관관계가 높은 것으로 보아 RMA 방법이 마이크로어레이 실험에서 발생하는 여러 오차를 제거하는 데 있어 비교적 효율적이다.

3.3. 각 방법들의 재현성

3.3.1. T-cell 결과 T-cell 데이터를 이용한 결과는 표 3.2와 같다. 5가지 유전자 탐색방법들에 대한 상관계수는 Pearson, Spearman방법 모두 높은 것으로 나타났다. 그러나 Pearson방법에 의한 Two-sample t-test방법의 상관계수는 0.60으로서 다른 방법들의 상관계수보다 약간 낮았다. 이 값은 결국 $9(={}_3C_2 \cdot {}_3C_2)$개의 상관계수에 대해 평균을 도출한 결과이고 그 값들 중 일부가 상당히 낮은 상관계수를 가졌기 때문이다. Two-sample t-test방법의 구조상 여기서는 변동을 고려하여 두 처리 집단-간 유의한 유전자를 탐색하는 것이고 변동의 영향을 많이 받아 통계량 자체가 너무 크고 때로는 너무 작게 도출되어 위와 같은 상관관계 값을 가지게 되는 것이다. 상위 20, 100개 유전자에 대한 일치성의 정도는 대체로 양호한 것으로 나타났으며, Two-sample t-test방법은 위와 같은 이유로 인해 상위 20개 유전자에 대한 일치성의 정도는 낮게 나타났다. SAM방법은 Two-sample t-test와 유사한 형태를 지닌 방법이지만 수정인자를 이용한 분모의 안정화를 통해 좀 더 높은 상관계수를 도출하였다.

표 **3.2.** T-cell Data

방법		상관계수			
		Pearson		Spearman	
MAS5	Fold Change	0.95	(0.01)	0.93	(0.01)
	Two Sample T	0.60	(0.16)	0.94	(0.01)
	SAM	0.96	(0.00)	0.94	(0.00)
	Wilcoxon	0.94	(0.00)	0.92	(0.01)
	LPE	0.96	(0.00)	0.86	(0.01)
RMA	Fold Change	0.99	(0.00)	0.98	(0.01)
	Two Sample T	0.50	(0.13)	0.95	(0.01)
	SAM	0.96	(0.01)	0.97	(0.01)
	Wilcoxon	0.96	(0.01)	0.94	(0.02)
	LPE	0.97	(0.02)	0.92	(0.02)

방법	유전자		평균	(표준오차)
MAS5	FoldChange	Top20	17.10	(0.35)
		Top100	80.00	(1.01)
	TwoSampleT	Top20	2.90	(0.48)
		Top100	34.80	(1.88)
	SAM	Top20	16.70	(0.29)
		Top100	83.40	(0.60)
	Wilcoxon	Top20	17.90	(0.63)
		Top100	82.30	(1.39)
	LPE	Top20	17.40	(0.65)
		Top100	83.40	(0.94)
RMA	FoldChange	Top20	18.80	(0.15)
		Top100	94.10	(0.59)
	TwoSampleT	Top20	3.60	(0.56)
		Top100	34.30	(1.55)
	SAM	Top20	11.40	(0.63)
		Top100	61.20	(2.47)
	Wilcoxon	Top20	15.00	(0.83)
		Top100	81.90	(2.84)
	LPE	Top20	15.00	(0.87)
		Top100	81.40	(4.40)

3.3.2. Prostate cancer 결과 Prostate cancer 데이터를 이용한 결과는 표 3.3과 같다. Prostate(전립선)데이터 에서는 한 그룹에 포함되는 표본의 수가 증가할수록 상관계수는 증가한다는 것을 알 수 있다. 그러나 LPE방법에 대한 상관계수가 다른 방법의 상관계수보다 낮음으로서 이 데이터는 LPE방법에 의해 잘 설명되고 있지 않다고 할 수 있다. 또한 상위 20, 100개 유전자의 일치성 정도는 표본의 수가 증가할수록 MAS5 변환방법에 의한 일치성이 좋음을 알 수 있다.

3.3.3. 기타 데이터 결과 HIV 데이터의 경우 LPE방법에 의한 상관계수가 낮으며 상위 20, 100개 유전자의 일치성은 SAM과 LPE방법이 가장 좋다. Clozapine데이터의 경우는 상위 20, 100개 유전자의 일치성 측면에서도 Prostate나 HIV데이터의 구조와 유사하다. Malaria데이터는 MAS5 표준화 방법을 이용한 Wilcoxon rank-sum test 방법의 상관계수는 Spearman 방법에서 상당히 낮은 상관계수를 가지며 상위 20, 100개 유전자의 일치성에서도 마찬가지로 낮은 값을 가진다. Head and

표 **3.3.** Prostate Data

		10 *vs.* 2				10 *vs.* 3				10 *vs.* 5			
		Pearson		Spearman		Pearson		Spearman		Pearson		Spearman	
MAS5	F	0.47	(0.01)	0.46	(0.01)	0.58	(0.01)	0.56	(0.01)	0.75	(0.01)	0.72	(0.01)
	T	0.28	(0.01)	0.46	(0.02)	0.55	(0.01)	0.58	(0.01)	0.74	(0.01)	0.74	(0.01)
	S	0.48	(0.02)	0.48	(0.02)	0.59	(0.01)	0.58	(0.01)	0.75	(0.01)	0.74	(0.01)
	W	0.47	(0.02)	0.46	(0.02)	0.58	(0.01)	0.57	(0.01)	0.75	(0.01)	0.73	(0.01)
	L	0.33	(0.02)	0.23	(0.01)	0.42	(0.01)	0.32	(0.01)	0.57	(0.01)	0.46	(0.01)
RMA	F	0.50	(0.04)	0.45	(0.04)	0.62	(0.03)	0.55	(0.03)	0.78	(0.01)	0.71	(0.02)
	T	0.27	(0.02)	0.40	(0.02)	0.53	(0.02)	0.52	(0.02)	0.71	(0.01)	0.67	(0.01)
	S	0.24	(0.03)	0.16	(0.03)	0.58	(0.02)	0.44	(0.05)	0.73	(0.01)	0.65	(0.02)
	W	0.42	(0.02)	0.39	(0.02)	0.53	(0.01)	0.50	(0.02)	0.69	(0.01)	0.65	(0.01)
	L	0.37	(0.05)	0.28	(0.04)	0.48	(0.03)	0.38	(0.03)	0.63	(0.02)	0.53	(0.02)

			10 *vs.* 2		10 *vs.* 3		10 *vs.* 5	
			평균	표준오차	평균	표준오차	평균	표준오차
MAS5	F	Top20	2.50	(0.43)	4.15	(0.52)	7.40	(0.37)
		Top100	16.20	(1.11)	23.60	(1.11)	37.60	(0.49)
	T	Top20	0.20	(0.09)	0.60	(0.18)	2.30	(0.37)
		Top100	3.45	(0.48)	8.75	(0.72)	23.85	(1.26)
	S	Top20	0.35	(0.11)	0.65	(0.20)	2.35	(0.26)
		Top100	4.70	(0.59)	8.75	(0.82)	24.50	(1.33)
	W	Top20	0.00	(0.00)	0.20	(0.12)	3.10	(0.27)
		Top100	4.50	(0.26)	8.30	(0.38)	21.05	(1.27)
	L	Top20	3.50	(0.46)	3.70	(0.26)	6.85	(0.33)
		Top100	19.25	(1.38)	24.35	(1.17)	35.00	(1.03)
RMA	F	Top20	4.20	(0.55)	6.55	(0.61)	10.65	(0.35)
		Top100	28.65	(2.46)	38.40	(2.52)	54.75	(0.65)
	T	Top20	0.60	(0.17)	2.80	(0.34)	6.40	(0.45)
		Top100	12.95	(1.75)	22.25	(1.77)	40.00	(1.69)
	S	Top20	0.25	(0.12)	1.90	(0.32)	3.90	(0.46)
		Top100	4.20	(0.62)	15.50	(1.17)	26.70	(2.16)
	W	Top20	0.60	(0.17)	1.95	(0.30)	6.40	(0.30)
		Top100	12.15	(1.60)	20.05	(1.81)	38.70	(1.81)
	L	Top20	3.00	(0.32)	4.35	(0.36)	5.60	(0.47)
		Top100	16.35	(1.64)	22.20	(1.24)	36.05	(1.30)

Neck Cancer(두경부암)데이터는 LPE방법의 상관계수가 다른 방법에 비해 낮으며 상위 20, 100개 유전자의 일치성 측면에서는 MAS5변환법의 Fold change방법이 좋게 나타났다. 이러한 결과의 차이는 각 데이터의 특성에 의존한다. 위에 제시되지 않은 나머지 데이터의 결과표 및 자세한 설명은 홈페이지(http://www.korea.ac.kr/˜stat2242/pub/)에 제시되어 있다.

4. 마이크로어레이 실험을 위한 표본 수 산정

4.1. 표본 수 산정 방법 및 실제 데이터에 적용

다중가설검정에 있어서 Jung (2005), Shao와 Tseng (2007)의 FDR을 고려한 표본 수 산정 방법에 대한 연구가 있었다. 그러나 효과의 크기(effect size)에 대해 좀 더 자세하고 현실적인 접근이 미비했던 것

은 사실이다. 본 논문에서는 기존 연구에 비해 다양한 효과크기와 세분화 된 FDR을 고려한 표본 수 산정방법을 제시하고, 또한 재현성의 정보를 고려하여 Jung (2005)의 표본 수 산정방법을 개선함으로써 마이크로어레이 실험을 실시하는 연구자에게 필요한 최소한의 표본 수에 대한 가이드라인을 제공하고자 한다.

4.1.1. 표본 수 산정 방법 Jung (2005)은 효과의 크기가 일정할 때 표본 수 산정에 필요한 여러 모수들과 식을 다음과 같이 제시하였다.

$$
\begin{aligned}
f &= \text{FDR level}\\
r_1 &= \sharp \text{ of true rejection}\\
a_k &= \text{allocation proportion for group } k\\
m_1 &= \text{number of genes expecting differentially expressed}\\
\delta_j &= \text{effect sizes for prognostic genes}\\
M_1 &= \text{the set of genes for which the alternative hypotheses are true}
\end{aligned}
$$

$$
n = \left[\frac{(z_{\alpha^*} + z_{\beta^*})^2}{a_1 a_2 \delta^2}\right] + 1, \qquad \alpha^* = \frac{r_1 f}{m_0(1-f)}, \qquad \beta^* = 1 - \frac{r_1}{m_1}.
$$

또한 효과의 크기가 다양할 경우 다음의 식이 0을 만족하는 n을 찾는 방법을 제시하였다. 아래의 식은 이분법(Bisection method)을 이용하여 해를 구한다.

$$
h(n) = \sum_{j \in M_1} \bar{\Phi}(z_{\alpha^*} - \delta_j \sqrt{n a_1 a_2}) - r_1.
$$

4.1.2. 표본 수 산정 방법의 실제 데이터에 적용 결과 본 논문에서는 Jung (2005)에서 제시한 몇몇 모수들에 대해 좀 더 현실적인 측면을 고려하여 실험에 필요한 표본 수를 산출하였다. 효과의 크기가 일정할 때와 일정하지 않을 때 표본 수를 산정한 결과파일은 홈페이지(http://www.korea.ac.kr/~stat2242/ pub/)에 제시되어 있다.

효과의 크기가 일정할 때 표본 수 산정결과를 보면, 효과의 크기가 증가할수록, FDR과 m_1이 증가할수록 실험에 필요한 표본 수는 감소한다는 것을 알 수 있다. 또한 효과의 크기가 다양할 때 3가지 경우의 효과의 크기를 고려한 표본 수 산정결과에서도 위에서 언급된 특정 모수의 변화에 의해 표본 수가 달라진다는 것을 알 수 있다. 그러므로 실제 마이크로어레이 실험을 위한 표본 수 산정에서는 효과의 크기가 다양할 때의 산정방법이 좀 더 현실적이라 할 수 있다.

4.2. 각 실험별 필요한 표본 수 산정

4.1장에서는 Jung (2005)이 제시한 부분에서 좀 더 세분화 하여 표본 수를 산정하였다. 그러나 효과의 크기를 고려하는 부분에 있어서 오래 전부터 사용해 오던 효과크기의 단위(0.5, 1, 1.5, 2)만을 적용했다는 점으로 볼 때 각 실험 고유의 특성을 무시한 경향이 있다. 4.2절에서 제시할 부분은 각 실험 별로 다르게 효과의 크기를 측정하여 표본 수를 산정하는 방법이다. 우선 효과의 크기와 추정된 효과의 크기는 아래와 같다.

$$
\delta_j = \frac{E(X_j) - E(Y_j)}{\sigma_j}, \quad j = 1, \ldots, m, \qquad \hat{\delta}_j = \frac{\bar{X}_j - \bar{Y}_j}{s_j}, \quad j = 1, \ldots, m.
$$

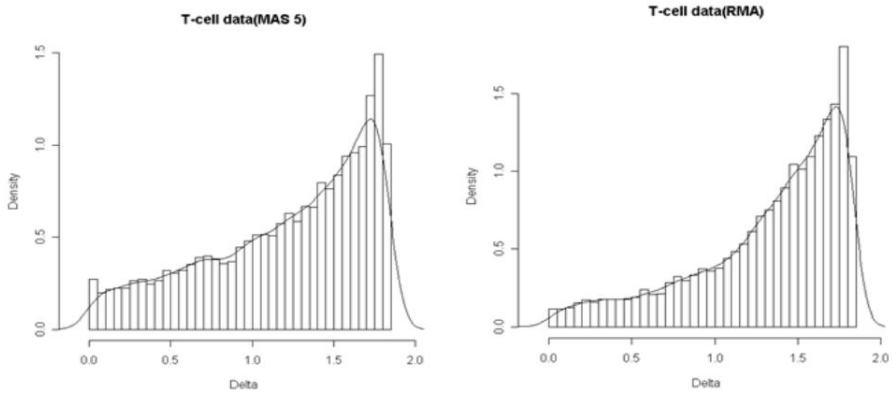

그림 **4.1.** T-cell 데이터(MAS5, RMA)의 효과의 크기 히스토그램

위와 같이 효과의 크기를 각 유전자마다 모두 구하면 m개의 $\hat{\delta}_j$가 산출되고 이에 대한 2, 3, 4 분위수를 구한다음 이들을 효과의 크기로서 가정한 후 표본 수를 산출하는 방법이다. 이러한 방법으로 효과의 크기를 구하는 이유는 마이크로어레이 실험에 필요한 표본 수 산정과정에 있어서 이 방법을 통해 특정 질병 또는 암이 가지고 있는 각 고유의 특성을 무시하지 않고 최대한 고려할 수 있기 때문이다.

앞 절에서 언급된 유전자 탐색방법(Fold change, Two-sample t-test, SAM, Wilcoxon rank-sum test, LPE)들 각각에 대해 상관계수를 제시하고 상위 유전자의 일치성을 확인한 것은 각 질병이 가지고 있는 고유의 특성을 확인한 일련의 과정이다. 이 방법은 본 논문에서 제시된 데이터와 동일한 질병에 대해 실험을 하기 위해 표본 수를 산정할 때 가이드라인으로써의 역할을 할 것으로 본다. 위의 방법을 이용하여 각 실험 별로 필요한 표본의 크기를 산출한 결과는 아래에 제시되어 있다. 표본의 크기를 산출할 시 4.2절에서 제시된 부분들을 동일하게 고려하였으며, 발현 값에 대한 표준화 방법(MAS5, RMA)각각에 대해서도 표본 수 산정결과를 제시하였다.

4.2.1. T-cell Data 결과 그림 4.1은 T-cell 데이터 발현값에 대한 표준화 방법(MAS5, RMA)각각에 대한 효과의 크기 히스토그램이다. T-cell 데이터에 대해 발현값에 대한 표준화 방법(MAS5, RMA)각각에 대해 효과의 크기를 살펴보면 RMA에 의한 표준화 방법이 MAS5 표준화 방법에 비해 효과의 크기가 대체로 크다는 것을 알 수 있다. 이를 토대로 표본 수를 산정할 때 3개의 사분위수를 효과의 크기로 이용한 표본 수 산정 결과는 표 4.1과 같다. 표 4.1를 보면 RMA 변환법을 이용한 T-cell 데이터에 대해 필요한 표본수가 MAS5변환법에 비해 작다는 것을 알 수 있다. 이 데이터에 대한 마이크로어레이 실험을 실시하고자 할 때 연구자가 발현값에 대한 표준화 방법으로 MAS5를 선택하고 각 그룹에 할당되는 표본의 비율을 동일하게 한 후($a_1 = 0.5$) 검정력을 80%(r_1/m_1), FDR = 1%로 설정하길 원한다면 필요한 표본수는 36이 되고 각 그룹에 18개씩 할당된다.

4.2.2. Prostate cancer 결과 그림 4.2를 통해 Prostate 데이터에서 각 표준화 방법 별로 효과의 크기를 보면 MAS5는 0에서 1사이에, RMA는 0에서 0.5사이에 효과의 크기가 대부분 분포하고 있으며 이는 T-cell 데이터와는 다른 양상이다. 표 4.2는 Prostate 데이터에 대한 표본 수 산정 결과표이다. MAS5는 검정력이 약 70%, RMA는 약 40%를 기준으로 실험에 필요한 표본 수가 급속하게 증가한다는 것을 알 수 있다.

표 **4.1.** T-cell 데이터를 이용한 표본 수 산정 결과표

Method	a_1	m_1	r_1/m_1	FDR					
				1%	2%	3%	5%	10%	20%
MASS	0.5	125	0.1	14	12	12	11	9	8
			0.3	19	17	16	15	13	11
			0.5	24	22	21	19	17	15
			0.7	31	28	27	25	22	20
			0.9	45	42	40	37	34	30
		624	0.1	11	9	9	8	6	5
			0.3	15	13	12	11	9	7
			0.5	19	17	16	15	12	10
			0.7	25	23	21	19	17	14
			0.9	37	34	32	29	26	22
	0.7	125	0.1	16	15	14	12	11	9
			0.3	22	20	19	18	16	13
			0.5	29	26	25	23	20	17
			0.7	37	34	32	30	27	23
			0.9	53	49	47	44	40	35
		624	0.1	12	11	10	9	7	6
			0.3	18	16	14	13	11	9
			0.5	23	20	19	17	15	12
			0.7	30	27	25	23	20	16
			0.9	44	40	38	35	31	26
RMA	0.5	125	0.1	13	12	11	10	9	7
			0.3	18	16	15	14	12	11
			0.5	23	21	20	18	16	14
			0.7	28	26	25	23	21	18
			0.9	40	37	35	33	30	27
		624	0.1	10	9	8	7	6	5
			0.3	14	13	12	10	9	7
			0.5	18	16	15	14	12	10
			0.7	23	21	20	18	15	13
			0.9	33	30	28	26	23	20
	0.7	125	0.1	16	14	13	12	10	9
			0.3	21	19	18	17	15	13
			0.5	27	25	23	21	19	16
			0.7	34	31	30	27	25	21
			0.9	47	44	42	39	36	32
		624	0.1	12	10	10	9	7	6
			0.3	17	15	14	12	10	8
			0.5	21	19	18	16	14	11
			0.7	27	25	23	21	18	15
			0.9	39	36	34	31	27	23

4.2.3. 기타 데이터 결과 Clozapind데이터와 Malaria 데이터의 경우 검정력이 약 40%를 기준으로 실험에 필요한 표본 수가 급속히 증가한다. Head and Neck Cancer데이터의 경우 RMA표준화 방법에 의해 도출된 효과의 크기로 표본 수를 산정하게 되면 그 효과의 크기가 상당히 작기 때문에 상당히 큰

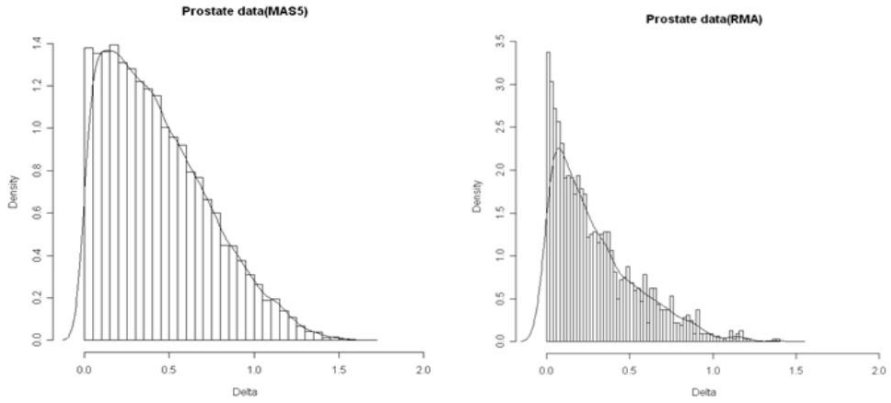

그림 4.2. Prostate 데이터(MAS5, RMA) 효과의 크기 히스토그램

표본 수가 산정되게 된다. MAS5표준화 방법에 의해 산정된 표본수는 검정력을 크게 설정할수록 큰 표본수가 필요하게 되며 이는 다른 데이터의 결과와도 비슷한 양상이다. 자세한 결과표 및 설명은 홈페이지(http://www.korea.ac.kr/˜stat2242/pub/)에 제시되어 있다.

5. 결론 및 토의

본 논문에서는 유의한 유전자 탐색 방법 중 5가지 검정방법(Fold change, Two-sample t-test, Wilcoxon rank-sum test, SAM, LPE)을 이용하여 6가지 실제 마이크로어레이 실험 자료에 대한 재현성을 비교하였다. 또한, 특정 질병에 관련된 마이크로어레이 실험을 하고자 하는 연구자에게 현실적인 정보를 이용하여 실험에 필요한 표본 수 산정에 대한 가이드라인을 제공하고자 하였다. 이를 위해 데이터의 재현성이라는 정보를 이용하여 이전부터 이슈가 되어 온 효과의 크기를 각 질병 별로 객관화 시키고자 하였다.

3장에서는 마이크로어레이 데이터별로 재현성을 측정해 보았는데, T-cell 데이터의 경우 Two-sample t-test를 제외한 다른 방법들은 비교적 높은 재현성을 가지는 것을 에서 확인할 수 있었으며 마이크로어레이 실험에서 발생하는 여러 오차를 제거하는데 있어서는 RMA 방법이 MAS5보다 비교적 효율적인 것으로 여겨진다. Prostate cancer 데이터의 경우 마찬가지로 RMA방법이 MAS5에 비해 효율적이며 5가지 유전자 탐색 방법이 비교적 비슷한 재현성을 보였다. 또한 재현성을 비교할 때 처리집단 별로 비복원 추출한 표본의 수가 커질수록 재현성이 더 높은 것을 확인할 수 있었다.

본 논문에서 언급된 실험들은 모수, 비모수적 방법을 통해 각기 다른 특징을 가지고 있다는 것을 확인할 수 있었으며 마이크로어레이 실험에 필요한 표본 수 산정에 있어 모든 질병에 동일한 기준과 방법으로 접근하는 것은 잘못되었다는 것을 본 논문의 결과에서 보여주고자 하였다. 이를 위해 4.1장에서는 Jung (2005)에서 제시한 모수들의 현실적인 측면을 고려하기 위해 효과의 크기가 일정할 때와 일정하지 않을 때 모수들을 변화시켜 나가면서 표본수를 산정해 보았다. 그 결과 효과의 크기, FDR, 이 증가할수록 실험에 필요한 표본 수가 감소하는 것을 알 수 있었다. 또한 3장에서 재현성을 비교한 데이터의 표본수를 산정하기 위하여 4.2장에서는 실험별로 효과의 크기를 추정한 후 사분위수를 이용하여 이 값들을 효과의 크기로 가정하는 방법을 이용하였다. 이 방법을 통하여 각 마이크로데이터의 고유한 특성을 효과의 크기에 반영할 수 있었다.

본 연구에서는 비교집단이 두 개일 경우의 마이크로어레이 데이터만을 이용하였다. 그러므로 비교 집단이 두 개 이상일 경우와 짝지어진 표본일 경우에 대해서도 추후에 연구가 지속되어야 한다. 또한 표본

표 **4.2.** Prostate 데이터를 이용한 표본 수 산정 결과표

Method	a_1	m_1	r_1/m_1	FDR					
				1%	2%	3%	5%	10%	20%
MASS	0.5	223	0.1	14	12	12	11	9	8
			0.3	19	17	16	15	13	11
			0.5	24	22	21	19	17	15
			0.7	31	28	27	25	22	20
			0.9	45	42	40	37	34	30
		1114	0.1	11	9	9	8	6	5
			0.3	15	13	12	11	9	7
			0.5	19	17	16	15	12	10
			0.7	25	23	21	19	17	14
			0.9	37	34	32	29	26	22
	0.7	223	0.1	16	15	14	12	11	9
			0.3	22	20	19	18	16	13
			0.5	29	26	25	23	20	17
			0.7	37	34	32	30	27	23
			0.9	53	49	47	44	40	35
		1114	0.1	12	11	10	9	7	6
			0.3	18	16	14	13	11	9
			0.5	23	20	19	17	15	12
			0.7	30	27	25	23	20	16
			0.9	44	40	38	35	31	26
RMA	0.5	223	0.1	13	12	11	10	9	7
			0.3	18	16	15	14	12	11
			0.5	23	21	20	18	16	14
			0.7	28	26	25	23	21	18
			0.9	40	37	35	33	30	27
		1114	0.1	10	9	8	7	6	5
			0.3	14	13	12	10	9	7
			0.5	18	16	15	14	12	10
			0.7	23	21	20	18	15	13
			0.9	33	30	28	26	23	20
	0.7	223	0.1	16	14	13	12	10	9
			0.3	21	19	18	17	15	13
			0.5	27	25	23	21	19	16
			0.7	34	31	30	27	25	21
			0.9	47	44	42	39	36	32
		1114	0.1	12	10	10	9	7	6
			0.3	17	15	14	12	10	8
			0.5	21	19	18	16	14	11
			0.7	27	25	23	21	18	15
			0.9	39	36	34	31	27	23

수 선정에 있어서 사분위수를 이용한 방법 이외에 좀 더 다양한 효과의 크기를 두루 반영할 수 있는 방법이 필요할 것이다. 마이크로어레이 데이터라 함은 본 논문에서는 Affymetrix사의 GeneChip®을 지칭하는 것으로 하였다. Affymetrix사의 GeneChip®은 반도체 제조공정과 비슷한 공정 과정을 거쳐 생

산되기 때문에 정확도 면에서 월등하다고 할 수 있으나 고비용의 문제가 존재하기 때문에 이 칩을 이용한 마이크로어레이 실험을 하기가 어려운 실정이다. 추후에 이 GeneChip®의 비용이 저렴해 진다면 이 칩을 이용한 실험이 보편화 될 가능성이 높기 때문에 본 논문의 높은 활용도 또한 기대할 수 있다. 마지막으로, 좀 더 많은 유전자를 집적할 수 있는 기술개발이 이루어진다면 그에 맞는 방법론도 필요할 것으로 생각된다.

참고문헌

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.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics*, **29**, 1165–1188.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science*, **18**, 71–103.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4**, 249–264.

Jain, N., Cho, H. J., O'Connell, M. and Lee, J. K. (2005). Rank-invariant resampling based estimation of false discovery rate for analysis of small sample microarray data, *BMC Bioinformatics*, **6**, 187.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M. and Lee, J. K. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays, *Bioinformatics*, **19**, 1945–1951.

Jung, S. H. (2005). Sample size for FDR-control in microarray data analysis, *Bioinformatics*, **21**, 3097–3104.

Shao, Y. and Tseng, C. H. (2007). Sample size calculation with dependence adjustment for FDR-control in microarray studies, *Statistics in Medicine*, **26**, 4219–4237.

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

Reproducibility and Sample Size in High-Dimensional Data

Won Seok Seo[1] · Jeea Choi[2] · Hyeong Chul Jeong[3] · HyungJun Cho[4]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University
[3]Department of Applied Statistics, University of Suwon
[4]Department of Statistics, Korea University

(Received August 2010; accepted August 2010)

Abstract

A number of methods have been developed to determine sample sizes in clinical trial, and most clinical trial organizations determine sample sizes based on the methods. In contrast, determining sufficient sample sizes needed for experiments using microarray chips is unsatisfactory and not widely in use. In this paper, our objective is to provide a guideline in determining sample sizes, utilizing reproducibility of real microarray data. In the reproducibility comparison, five methods for discovering differential expression are used: Fold change, Two-sample t-test, Wilcoxon rank-sum test, SAM, and LPE. In order to standardize gene expression values, both MAS5 and RMA methods are considered. According to the number of repetitions, the upper 20 and 100 gene accordances are also compared. In determining sample sizes, more realistic information can be added to the existing method because of our proposed approach.

Keywords: Microarray, reproducibility, sample size, effect size.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2009-0087564).
[4]Corresponding author: Associate Professor, Department of Statistics, Korea University, Anam-Dong 5-Ga, Seongbuk-Gu, Seoul 136-701, Korea. E-mail: hj4cho@korea.ac.kr