Cancer Research and Treatment 2003;35(6):533-540

cDNA Microarray Experiment: Design Issues in Early Stage and the Need of Normalization

Byung Soo Kim, Ph.D.[1], **Sunho Lee,** Ph.D.[2], **Sun Young Rha,** M.D., Ph.D.[3,4] and **Hyun Cheol Chung**, M.D., Ph.D.[3,4]

[1]Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea; [2]Department of Applied Mathematics, Sejong University, Seoul 143-747, Korea; [3]Cancer Metastasis Research Center, College of Medicine, Yonsei University, Seoul 120-752, Korea; [4]Brain Korea 21 Project for Medical Science, College of Medicine, Yonsei University, Seoul 120-752, Korea

Purpose: **The cDNA microarray has become a useful tool for observing the expression of thousands of genes simultaneously. However, obtaining good quality microarray data is not easy due to the inherent noise at various stages of the experiment. Therefore, it is essential to understand the source of the variation in the microarray experiment and its size as an initial step of the data analyses.**

Materials and Methods: **The total RNA extracted from HT-1080 fibrosarcoma and normal rat tissues were hybridized to the cDNA microarrays with 0.5 K human and 5 K rat genes, respectively. A homotypic reaction and dye swap experiments were used to identify the sources of the variation.**

Results: **The relative fluorescent intensities of the microarray, if unnormalized, have a large variation, particularly in the lower intensity region. The distribution of the log intensity ratios also exhibit some departure from a band around zero, which is the distribution pattern expected when the majority of genes in the microarray are not regulated. Normalization of the log ratios is usually required as a means of preprocessing the data. We claim that a within-print tip group, an intensity-dependent normalization through a loess fit adjustment will be useful for this purpose, particularly in the initial stages of the microarray experiment.**

Conclusion: **For proper data analysis, an understanding the source of the variation and preprocessing of data with a suitable normalization method will be important. It is important to have an interactive cooperation between a researcher and a statistician from the early stages of the study design and to the final stages of data analysis.** ***(Cancer Research and Treatment 2003;35:533-540)***

Key Words: **cDNA microarray, Homotypic experiment**

서 론

생물학의 발전과 함께 생물현상은 수많은 유전자들의 역동적인 작용의 결과임이 제시되고 있다. DNA 마이크로어레이는 수천~수만 개 유전자의 발현 수준을 동시에 포괄적으로 조사할 수 있도록 해주는 고효율 분석 도구로 이미 자리를 잡고 있다(1). 발암기전이나 예후를 예측하는 암 관련 연구(2~4)뿐 아니라, 생태학(5), 독성학(6) 등 많은 분야에서 다양하게 활용되고 있다. DNA 마이크로어레이를 이용한 암 관련 연구로는 1) 종양의 유전학적 성상(molecular signature)과 생물학적 특성(tumor biology)을 이해하는 기본 자료를 제공하며, 2) 암의 정확한 진단 및 아형의 구분과 발견을 용이하게 한다. 3) 또한 암의 진행과 예후에 관련된 유전학적 정보를 근거로 하여 예후를 예측하고 치료의 기준을 선정하는 임상에의 응용이 가능하게 된다. 4) 이와 같이 발굴된 분자생물학적 표지자들의 기능이 밝혀지면서 이들을 치료제 개발의 목표로도 이용할 수 있는 다양한 분야가 있다(7~12).

그러나 이와 같은 많은 정보를 주고, 다양한 연구 분야에 이용될 수 있는 장점이 있는 반면, 그 결과를 이해하고 정확히 분석하여 유용한 정보를 얻기 위해서는 많은 기술적, 생물학적, 및 통계학적 문제점을 해결하여야 한다. 마이크로어레이 실험도 여타 생물학 실험과 마찬가지로 실험 설계가 우선되어야 하는데, 실험 설계는 실험 목적에 의존적일

Correspondence: Byung Soo Kim, Department of Applied Statistics, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul 120-749, Korea. (Tel) 82-2-2123-4541, (Fax) 82-2-313-5331, (E-mail) bskim@yonsei.ac.kr

Received October 22, 2003, Accepted December 24, 2003
The study of BS Kim was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (02-PJ1-PG3-10411-0003). The study of SY Rha was supported by a grant of Yonsei University 2001-1-024.

뿐 아니라, 분석 방법을 염두에 두고 결정되어야 한다. 또한 마이크로어레이는 실험의 속성상 여러 단계를 거치며 매 단계마다 실험 오차가 발생할 여지가 많고, 일반적으로 오차가 많은 실험이라고 할 수 있다. 여러 단계에서 누적된 오차는 통계 분석을 어렵게 하는 한 요인이 된다. 그러나 이보다 더 통계학자들을 "당혹"스럽게 하는 것은 수천 수만 개의 유전자중에서 불과 몇 개를 선별하여야 하는데 분석 대상의 관찰치를 구성하는 마이크로어레이는 불과 수십 개에 불과한 이른바 "큰 p, 작은 n"의 문제를 구성하고 있다는 점이다. 여기서 p는 설명변수(유전자)의 개수이고, n은 관찰치(마이크로어레이)의 개수를 나타내고 있다. 전통적으로 통계학자들은 "큰 n, 작은 p" 자료를 주로 다루어 왔으나 마이크로어레이 자료가 대표적인 "큰 p, 작은 n" 자료를 구성하고 있다. 따라서 마이크로어레이 실험 자료의 통계적 분석은 통계학자들에게도 새로운 도전이 되고 있다(13).

이 연구에서는 국내실험실에서 많이 사용하고 있는 cDNA 마이크로어레이에 국한하여 지난 2년간 필자들이 초기 실험과정에서 경험하였던 실험 설계와 표준화의 필요성에 대하여 개관하고, 실제 실험 자료를 사용하여 설명하고자 한다.

재료 및 방법

1) 세포주 배양 및 RNA 추출

ATCC (Rockville, MD)에서 구입한 HT-1080 fibrosarcoma 세포주를 사용하였다. 기본 배지인 minimum essential media (GIBCO, Grand Island, NY)에 10% 우태아 혈청(fetal bovine serum; Commonwealth Serum Laboratories, Melbourne, Australia), penicillin (100 U/ml; GIBCO)과 streptomycin (100μg/ml; GIBCO)을 첨가하여 사용하였다(14). 5% CO_2의 존재하에 37°C 항온 배양기에서 배양하며 세포가 배양용기의 95% 이상 자라게 되었을 때 세포를 수확하여 Trizol (InVitrogen Co., Carlsberg, CA)을 이용하여 total RNA를 추출하였다. Rat chip 실험을 위하여 정상 rat의 간, 폐, 심장, 비장, 신장, 위장 등의 장기를 적출하여 -80°C에 보관하였다. 조직의 경우는 액체 질소하에 조직이 얼어있는 상태에서 막자사발을 이용하여 분쇄 후 세포주와 동일한 방법으로 RNA를 사용하였다. Total RNA의 양과 질은 Gene spec I spectrophotometer (Hitachi, Inc., Tokyo, Japan), 전기영동 및 Bioanalyzer (Agilent, Inc., Palo Alto, CA)를 이용하여 확인하였다.

2) cDNA microarray 실험

cDNA microarray는 480개의 인간 유전자가 점적되어있는 cDNA microarray (CMRC-GT, Seoul, Korea)와 5,000개의 rat 유전자가 점적되어 있는 rat chip (Genomic Tree Co., Daejeon, Korea)를 사용하였다. Microarray 실험은 연세의대 암전이연구센터에서 확립한 protocol에 따라 시행하였다. 간단히 설명하면, 50μg의 total RNA를 역전사과정을 통하여 대조군에는 Cy-3를, 실험군에는 Cy-5를 표지시키는 probe를 제작한 후, 칩위에 동시에 보합시켰다. Probe를 제작하기 위하여 400 units의 SuperScript II (GIBCO), 3 ml Cy5-dUTP (or Cy-3 dUTP), 1.5 ml의 dATP, dCTP 와 dGTP, 0.6 ml dTTP, 6μl의 5X first-strand buffer, 4 ug의 modified Oligo-dT primer를 total RNA와 같이 혼합하여 42°C에서 2시간 동안 반응시켰다. 반응이 끝나고 역전사가 되지 않은 nucleotide는 PCR Purification Kit (Qiagen, Valencia, CA)을 이용하여 제거하였다. 준비된 probe는 20μl의 1μg/μl Human Cot1 DNA (GIBCO), 2μl의 10μg/μl polyA RNA (Sigma, St. Louis, MO), 288μl의 1M TE buffer와 혼합한 뒤, Microcon-30 tube (Millipore, Bedford, MA)을 사용하여 농축하였다. 48μl로 농축된 probe를 10.2μl의 20 X SSC와 1.8μl의 10% SDS와 혼합한 후, 95°C에서 2분간 가열하고 13,000 r.p.m.으로 2분간 원심분리 후 준비된 cDNA microarray 위에 보합작용을 시켰다. 칩은 비특이적 보합작용을 막기 위하여 10 mg/ml BSA, 3.5×SSC, 0.1% SDS solution 내에서 2시간 동안 반응을 시킨 뒤 사용하였다. 65°C에서 16시간 동안 보합작용을 시킨 후, 다음과 같이 4회 수세를 하였다(2×SSC 10분, 0.1% SDS 10분, 0.1×SSC 10분간 2회). 수세가 끝나고 600 r.p.m.으로 5분간 원심분리하여 말린 후 GenePix 4000B (Axon Ins., Foster City, CA) scanner를 이용하여 image를 얻고, 그 영상자료는GenePix Pro3.0 (Axon Inc.)를 이용하여 기본 자료로 전환되었다.

3) 상동 실험(Homotypic experiment)

같은 RNA 시료에 cy3와 cy5를 표지한 후 관찰한 실험을 1 : 1 상동실험 (homotypic experiement)이라 한다. cDNA 마이크로어레이 실험에서 여러 가지 변동원의 실체와 크기를 파악하고, 표준화에 대한 필요성을 알아보기 위해 0.5 K 마이크로어레이에 인간 섬유육종 세포주인 HT-1180RNA를 사용하여 1 : 1 상동 실험을 실시하였다. 또한 5 K의 rat 유전자가 점적되어있는 마이크로어레이를 이용하여 쥐의 정상 조직 pool을 이용하여 1 : 1 상동으로 보합실험을 시행하였다. 이 때 0.5 K 마이크로어레이는 4개의 프린트 팁을 사용하여 점적하였고, 5 K의 경우에는 16개의 프린트 팁을 사용하였다.

4) 복제 관찰 실험

복제관찰의 중요성을 확인하기 위하여 0.5 K 마이크로어

레이에 HT-1080 RNA를 이용하여 두 조의 실험을 시행하였다. 제1조 실험은 기술적 복제 관찰치 세 개를 둔 것으로, Total RNA를 세 개의 뱃치로 나누어서, 각 뱃치로 부터 cDNA를 역전사하였고, 각각 다른 날, 같은 조건하에서 보합시켰다. 제2조 실험에서는 반복측정에 가까운 수준으로, 동일 RNA를 한번의 역전사작용을 시킨 후, 이를 세 뱃치로 나누어서 제1조와 같은 조건하에서 같은 시간에 보합시켰다.

5) 영상자료 분석

cDNA 마이크로어레이 실험 자료 분석의 첫 단계로, 이후 계속하여 진행되는 data preprocessing과 실험의 원 목적인 특이 유전자 발현 검색, 군집 분석 등에 많은 영향을 미칠 수 있는 중요한 단계이다. 실험이 끝난 어레이는 스캐너에 의하여 각 스팟에서 적색과 녹색 형광 강도를 계측하고, 그 두 색의 형광 강도는 점적된 DNA와 두 시료가 각각 보합된 정도를 나타낸다. 한 개의 스팟은 200～400개의 픽셀(pixel)로 구성되고, 스캐너의 PMT (photomultiplier tube)가 픽셀 단위에서 방출된 광자를 전자로 바꾸고, 이어 A/D (analog to digital) 변환기에 의하여 전자는 0～65535 (2^{16}～1) 사이의 숫자로 전환한다. 이 때 대표값으로서는 극단치 등을 고려하여 평균보다는 중위수(median)가 선호된다. 즉, Cy-3dUTP에 의한 녹색염료의 강도는 G (green intensity)=Gfg (foreground)-Gbg (background)로, Cy-5dUTP에 의한 적색염료의 강도는 R (red intensity)=Rfg-Rbg로 나타낸다. 스캐너에 의하여 처리되는 영상자료의 분석은 각 스팟의 위치 확인, 전경과 배경 분할(segmentation of foreground and background) 및 강도 추출의 세 단계를 거치게 된다(15). 이 세 단계 모두에서 어떠한 기준을 사용하느냐에 따라 오차를 조절할 수 있고, 최근에는 마이크로어레이 제작의 기법이 발달하여 재현 가능한 양질의 칩과 자동화된 스캐너를 사용함으로써 여기서 발생되는 오차도 많이 감소하였다.

6) M-A plot 및 표준화

하나의 마이크로어레이에 p개의 유전자가 점적되어 있을 경우 p개의(R, G)를 관찰하게 되고, 이를 이차원 평면에 점산도로써 나타내면 Fig. 1의 R vs G plot가 된다. 그런데 R, G 각각에 로그를 취하면 강도가 낮은 지역에서의 R, G간의 행태를 더 자세하게 알 수 있다. 또한 로그로 변환한 상대 강도 M=log2 (R/G)과 로그로 변환한 두 염료 강도의 평균 A=log2 R+log2G/2를 이용한 M-A 그림은 log2 (R) vs log2 (G)을 시계 반대 방향으로 45도 틀어 놓은 M-A 그림이다. 스팟의 이상점 식별이나 상대강도가 강도 의존적으로 변하는 형태를 잘 나타내주므로 표준화를 설명하는 데 유용하여 본 논의뿐 아니라 많은 마이크로어레이 자료 분석에는 M-A 그림이 사용되고 있다(Fig. 1). 실험자료에 따른 표준화 방법간의 차이를 관찰하기 위하여 global normalization, intensity dependent normalization, within-print tip group intensity-dependent normalization을 시행하였다(15～18).

결 과

1) 마이크로어레이 실험방법과 발생 가능한 오차

마이크로어레이 실험의 결과에서 생길 수 있는 가장 기본적인 오차는 마이크로어레이 슬라이드의 제작에 있다. 즉, 많은 수의 유전자를 좁은 공간에 동일한 양의 유전자를 일정하게 점적하여야 하는데, 유전자 크기나 특성이 다른 수천 수만 개의 유전자 클론을 처리하여 오차가 없이 양질의

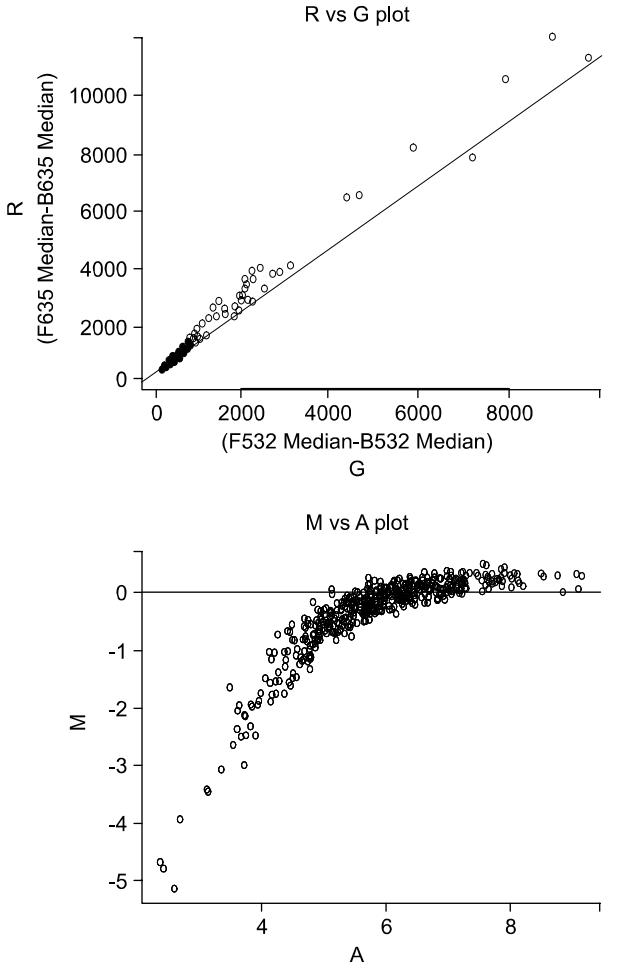

Fig. 1. Comparison of scatter plots between “R vs G” and “M-A” plots using a homotpyic experiment of 0.5 K microarray. The M-A plot is useful in identifying outliers and detecting intensity dependent patterns in the log ratios.

마이크로어레이를 제작하기는 매우 어렵다. 다행히 robotics의 발달로 최근 2년간 그와 같은 오차는 많이 줄어들고 있다.

마이크로어레이 실험의 기본 원칙은 두 가지 RNA를 다른 색으로 표지시켜(labeling) 많은 유전자가 점적되어 있는 유리 슬라이드 위에 보합작용을 시킨 뒤, 그 상대적인 발현을 비교 분석하는 것이다. 우선 추출된 RNA를 역전사(reverse transcription)과정을 거쳐서 표지하게 되는데, 이 때 사용되는 표지자로는 각각 녹색과 적색의 형광을 띠는 Cy-3dUTP와 Cy-5dUTP를 사용하게 된다. 이 과정에서 역전사된 RNA 양과 얼마만큼의 RNA가 역전사되는 지와 그 중 어느 정도의 RNA가 표지되는지를 정확하게 알 수는 없다. 또한 사용되는 두 가지 표지자의 크기와 표지 효율이 틀리므로 표지가 끝난 후 마이크로어레이 실험에 사용된 정확한 양을 예측하기 어렵다. 이와 같이 표지된 두 가지 시료는 슬라이드 위에 점적되어있는 유전자와 두 시료 간의 경쟁적인 보합작용을 하게 되는데, 그 정도 또한 정확히 수치화할 수는 없다. 보합작용이 끝난 슬라이드를 스캐너를 이용하여 image를 얻고 그 형광의 정도에 근거하여 그 결과를 수치화하는 과정에서도 유전자가 점적되어 있는 스팟과 주변 background의 결정을 어떻게 하느냐에 따라 많은 오차가 유발될 수 있다.

2) Data preprocessing: 표준화(normalization)

전술한 바와 같이 상대강도에 영향을 미치는 체계적 변동원은 다양하며 그중 대표적인 예가 1) 녹색 형광 염료가 적색 형광 염료보다 강도가 높은 경향이 있고, 2) 유전자를 점적하는 프린트 팁마다 점적되는 양이 다를 수 있으며, 3) 어레이의 어느 지역에 점적되었느냐에 따라 강도가 변하는 이른바 공간효과가 나타나고 있다. 이러한 체계적 효과는 비교 대상이 되는 RNA 시료 간 차이와는 무관하게 발생하므로, 자료에 적절한 변환을 취하여 제거하는 data preprocessing 과정이 필요하고, 그중 가장 대표적인 방법이 표준화 과정이다. 이와같이 실험의 목적이 되는 비교 대상 시료간의 차이 이외에 여타의 체계적 변동원을 제거하여 어레이간 비교를 가능케 하고자 함이 표준화의 목적이다.

대부분의 표준화 방법은 전체 p개의 유전자 중 특이 유전자의 비율이 작거나, 혹은 그 비율이 어느 정도 수준이라 하더라도 상향 조절된 유전자와 하향 조절된 유전자의 비율 및 크기가 유사하다는 가정 아래 시행한다. 따라서 전체 유전자를 대상으로 한 M-A 그림은 M=0을 중심으로 한 점산도를 나타낼 것이라는 것을 전제로 하고 있고 다양한 방법들이 제시되고 있다(16～18).

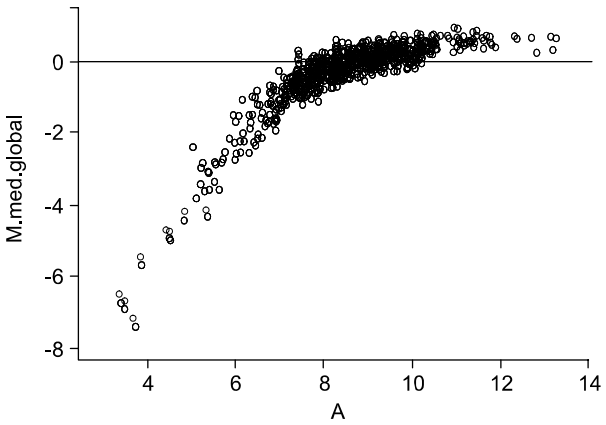

Fig. 2. Global normalization through the median adjustment using a 0.5 K microarray. Global methods adjust so that the new median is equal to zero. Black and orange circles denote M values before and after normalizations, respectively. Even after the normalization the nonlinear trend of the log ratios still exist and overall pattern is far from a band around M=0.

3) Global normalization

가장 단순한 방법으로서 p개의 M값의 중위수[median (M)]를 0으로 조정하는 방법이다. 이는 p개의 M값 각각에서 median (M)을 빼주면 된다. 이러한 전체 중위수 조정법은 Fig. 1과 같이 강도 의존적 추세를 보이는 경우, M-A 그림의 추세선만 위로 수직 이동한 결과(Fig. 2)가 되고, 애초에 의도하였던 M=0을 중심으로 한 일정한 간격의 띠 형태와는 거리가 있어 적절한 표준화 방법이 되지 못한다. 즉, Fig. 2의 자료에서 전체 중위수는 -0.251로 계산되는데 이를 480개 모든 M값에다 일률적으로 빼주면 Fig. 2에서 보듯이 결국 곡선의 추세선을 위로 수평 이동하는 결과가 되고, 애초에 의도하였던 M=0을 중심으로 한 일정한 간격의 띠 형태와는 거리가 있게 된다.

4) Intensity dependent normalization

두 번째 방법은 Fig. 1과 같이 강도가 낮은 지역에서는 M값이 낮게 나타나는 이른바 “강도 의존적” 현상이 나타날 때 사용하는 방법으로서, 비모수적 추세선을 적합하여 그 추세선을 M=0와 일치하도록 각각의 M값에 적당한 상수를 더하거나 빼주는 방법이다. 이러한 추세선은 lowess (locally weighted smoothing scatter plot) 적합법(18)을 통하여 구할 수 있으며, 실제로 Fig. 3에서 곡선으로 나타난 추세선이 lowess 적합을 통하여 얻어 졌다. 이 추세선을 M=0으로 끌어 당기거나 끌어 내리기 위하여 추세선 각 점에다 적당한 값을 더하거나 빼주면 된다.

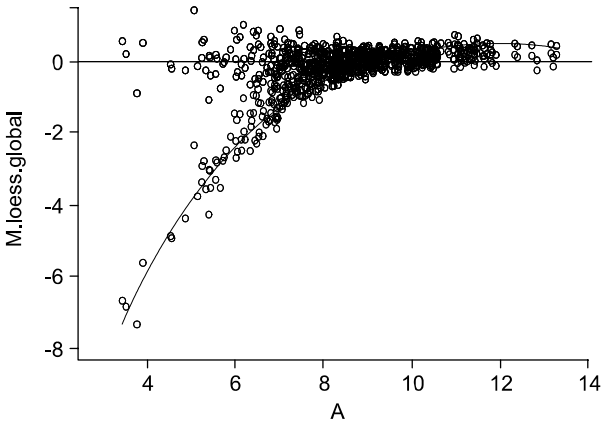

Fig. 3. Intensity-dependent normalization using lowess. First the lowess fitting curve (solid line in the black circles) is obtained using S+ software and then a constant is added or subtracted to each point so that the solid line is aligned at M=0. Black and orange circles denote M values before and after normalizations, respectively.

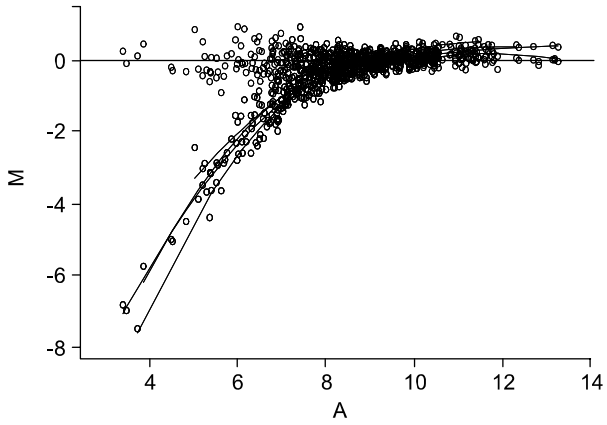

Fig. 4. Within-print tip group intensity-dependent lowess curves and normalization. Four lowess fit curves corresponding to four print tips show some inconsistency, which suggests the normalization of the separate intensity-dependent loess fit adjustment for each print tip group. Black and orange circles represent M values before and after normalizations, respectively. One may not find a big difference between this fig. and Fig. 3 Box plots of Fig. 5 detect the difference between these two normalization methods.

5) Intensity dependent, print tip normalization with loess method

세 번째 방법은 강도 의존적 lowess 적합 방법을 각 프린트 팁별로 적용하는 것으로 프린트 팁에 의한 공간 효과와 강도 의존적 추세를 동시에 보정하여 주는 방법이다. 프린트 팁에 의한 공간 효과와 강도 의존적 추세가 일반적 현상이고, Fig. 4에서도 4개의 프린트 팁별로 lowess 적합선을 추정하였는데 이 네 개가 모두 일치하는 것이 아니므로 프린트 팁별로 lowess 적합을 통한 보정이 바람직하다. 일부 실험자들은 강도가 “낮은” 스팟을 분석에서 제외하기도 한다. 가령 Fig. 4에서 “낮은” 강도의 스팟을 제거하여도 프린트 팁별로 추세선이 일치하지 않고, 또한 “중간” 강도 수준에서도 비선형 추세가 관찰되므로 프린트 팁별 lowess 보정을 통한 표준화가 전체 중위수 표준화보다 더 효율적이다. Fig. 3의 전체 강도 의존적 보정 방법을 사용하여 표준화한 결과와 Fig. 4의 프린트 팁별 강도 의존적 보정으로써 표준화를 실시함 M-A 그림은 큰 차이가 없어 보인다. 그러나 Fig. 5의 상자그림을 보면 표준화 이전, 전체 강도의존적 표준화 이후, 프린트 팁별 강도 의존적 표준화 이후 각각의 경우에서 M값의 프린트 팁별 상자 그림을 살펴보면 두 표준화 방법 간의 차이는 확연해진다. 전체 강도 의존적 방법의 결과는 전체 중위수를 0에 가깝게 만들어 놓았을 뿐 프린트 팁에 의한 공간 효과는 제거하지 못한 반면(Fig. 5B), 프린트 팁별 강도 의존적 방법은 4개 프린트 팁에 의해 구성된 4개 블록의 중위수를 모두 0에 일치시킴으로써 공간 효과도 함께 제거한 것을 알 수 있다(Fig. 5C). 현재 상용중인 스캐너의 사용 프로그램에서는 전체 중위수 보정에 의한 표준화 절차가 있을 뿐 프린트 팁별 loess 보정을 통한 표준화 방법은 제공되지 않는다. 프린트 팁별 loess 보정의 표준화 절차는 S+나 MATLAB 등과 같은 계산 전용 소프트웨어를 이용하여 쉽게 전산 코드화할 수 있다.

또한 5 K 랫드 유전자가 점적된 마이크로어레이를 사용하여 초기 실험을 하였을 때에도, Fig. 6에서 볼 수 있듯이 낮은 강도에서 곡선의 추세선이 관찰되었으며, 어레이의 오른쪽 끝 블록들이 다른 블록들과 구별되는 양상을 보이는 주변 효과도 관찰되었다. 이는 어레이 제작 시 프린트 팁이 4, 8, 12, 16에 해당하는 오른쪽 끝에 위치한 4개의 블록에서 다른 프린트 팁과 구별되는 양상을 보이기 때문이다. 이런 경우 프린트 팁별로 lowess 추세선이 일치하지 않으므로 표준화 방법으로 프린트 팁별 강도 의존적 방법을 사용하는 것이 바람직함을 확인할 수 있다.

6) 복제 관찰치의 종류와 의미

복제 관찰치(replicate)는 관심 대상의 성질을 여러 개의 독립적 개체로부터 관찰한 것을 말하며, 반복측정치(repeated measure)는 관심 대상의 성질을 한 개체에서 여러 번 관찰하는 것을 의미한다. 마이크로어레이 실험 변동원은 3개층으로 분류해 볼 수 있다. 첫 번째가 생물학적 변동으로 실제 종양이나 개인간 차이를 나타내는 것이고, 두 번째가 기술적 변동으로 RNA를 추출하거나, 표지, 보합과정에서

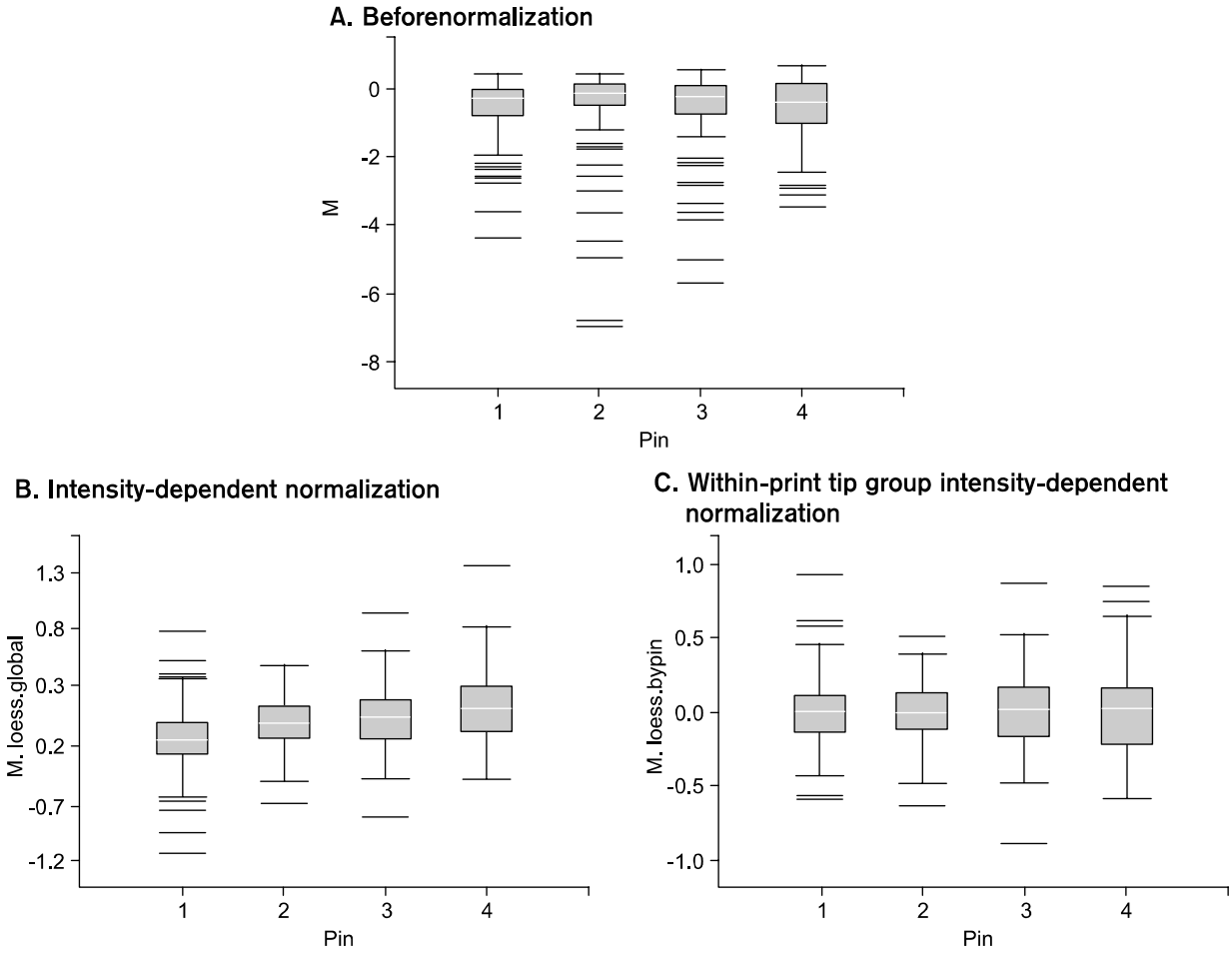

Fig. 5. Three box plots of log ratios (M) before normalization, intensity-dependent normalization and within-print tip group intensity-dependent normaization, respectively. Note that medians (white bars in boxes) are deviated from M=0 at (B), whereas these four medians are aligned at M=0 at (C).

생기는 변동이고, 세 번째가 계측오차로 먼지 등에 의하여 생길 수 있는 형광 신호의 해독과 관련된 변동이다. 생물학적 복제 관찰치와 기술적 복제 관찰치를 두는 것은 각각 개인 간 변동과 개인 내 변동을 고려하기 위함이고, 반복 관찰치를 두는 것은 계측오차를 줄이기 위함이다. 최근에는 기법의 발달로 계측오차로 인한 변동이 많이 줄어들어, 처음 두 단계의 변동이 주 관심사이며 그 두 가지 변동을 해결하는 기본적인 방법으로 복제실험과 복제 관찰치의 중요성이 많이 제시되고 있다. 초파리의 transcriptome의 변동원을 밝히는 실험에서 군당 생물학적 복제관찰치를 6개로 두었으며, 동 논문의 Fig. 2는 복제관찰치를 적게 두었을 때 생길 수 있는 오류의 가능성을 극명하게 보여주고 있다 (19).

이 연구에서는 동일 RNA를 나누어 실험한 1조와 역전사가 끝난 cDNA를 나누어 실험한 2조 실험 결과를 이용하여 복제관찰치의 재현성을 확인하였다. 각각의 세 개 복제 관찰치들 간의 재현성은 Table 1의 스피어만의 순위 상관 계수로써 나타낼 수 있으며 제1조 실험이 제2조 실험보다 재현성이 떨어지는 것을 확인할 수 있다. 따라서 기술적 복제 관찰치를 두는 것이 반복측정치 수준의 복제 관찰치를 두는 것보다 더 큰 변동을 유발함을 확인하였다. 마이크로어레이 자료가 정규분포에서 이탈된 분포를 보이기 때문에 재현성의 측도로는 피어슨 상관계수 대신에 스피어맨의 순

Table 1. Comparison of reproducibility measured in terms of Spearman's rank correlation among three replicates in two replicate sets

Array	Set 1		Set 2	
	2	3	2	3
1	0.846	0.837	0.90	0.903
2		0.829		0.863

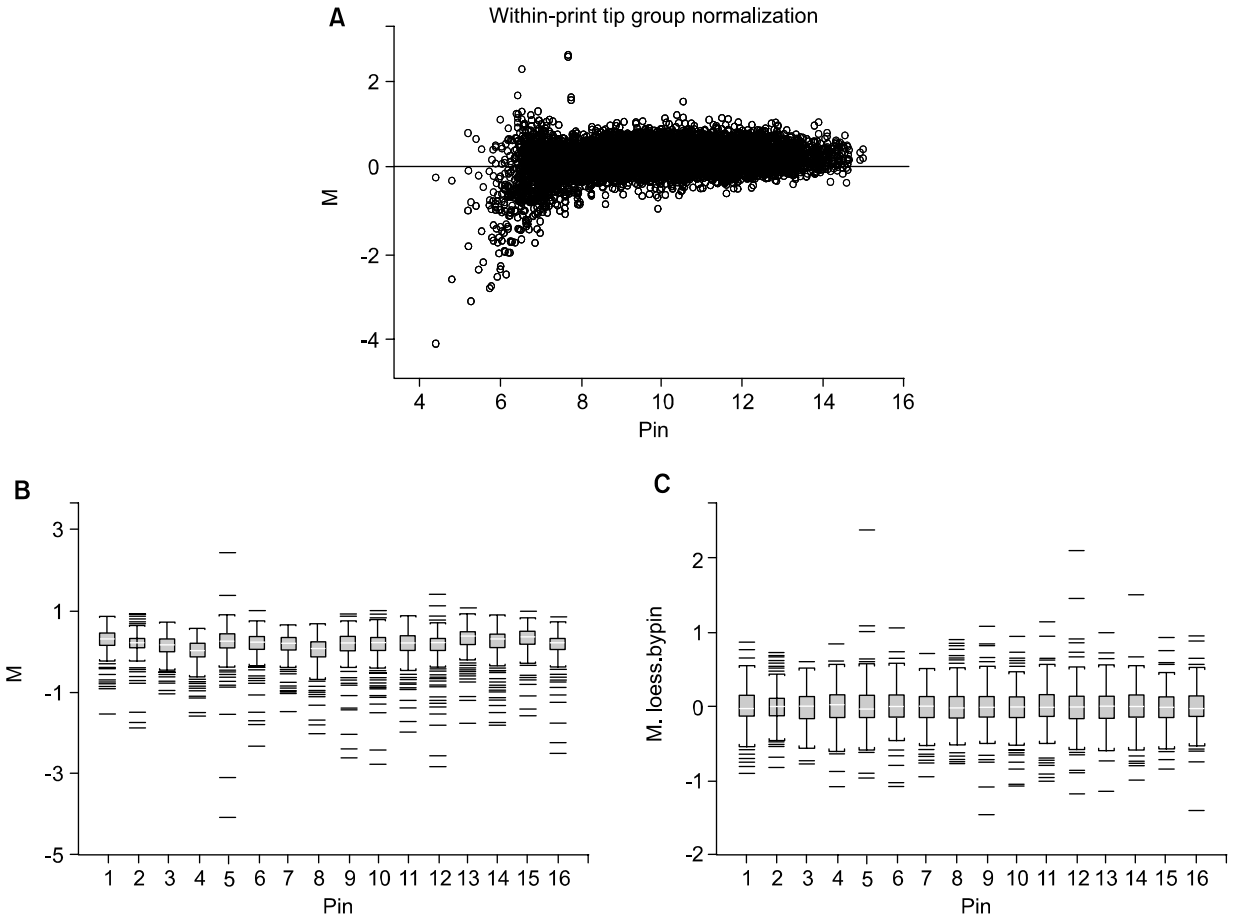

Fig. 6. M-A plots of 5 K rat cDNA microarray. (A). Black and orange circles denote M values before and after within-print tip group intensity-dependent normalizations. (B) and (C) represent box plots of M values with respect to 16 print tips before and after within-print tip group intensity-dependent normalizations. We may note in (B) that print tips corresponding to 4, 8, 12 and 16 yield lower M values relative to the other print tips. These four blocks are located at right edges of the array and thus this effect is referred to as the edge effect, a special type of the spacial effect.

위상관 계수를 사용하였다.

고 찰

cDNA 마이크로어레이 실험은 전술한 바와 같이 여러 단계를 거치며 각 과정마다 오차가 개입될 수 있으므로 우선 적절하게 설계된 초기 실험에서 각 단계별 오차의 크기를 파악하고, 그 유형과 의미를 파악하는 것이 중요하다. 우선 마이크로어레이 실험을 통하여 해결하려는 생물학적 질문을 통계적 질문으로 전환하는 과정이 있어야 한다. 이 과정에서 각 단계별 오차를 줄일 수 있는 실험설계 및 분석방법 선정이 필수적이므로 통계전문가와 미리 상의하고, 진행과정에서 지속적인 의견교환이 바람직하다.

복제관찰치의 크기를 어느 수준으로 하여야 하는가는 아직 미해결의 문제로 남아있다. 일반적으로 표본크기 n은 검색하고자 하는 차이(δ), 제1종 오류의 확률(α), 제2종 오류의 확률(β), 그리고 관찰치의 분산이 주어졌을 때 결정된다. 그러나, 한 마이크로어레이에 점적되어 있는 수천~수만개의 유전자들의 발현 정도의 분산이 일정하지 않고, 그 크기가 또한 알려져 있지 않으므로 현재까지의 이론으로 n을 계산하는 것은 어렵다(16,20,21).

결 론

cDNA 마이크로어레이 실험은 여러 단계의 과정을 거치고, 매 단계마다 오차가 개재될 수 있으므로, 재현성 있는 실험 결과를 얻기까지는 상당한 시간이 소요되는 실험이다. 초기 실험에서는 변동원과 그 크기를 파악하는 것이 매우 중요하고, 이를 위하여 1 : 1 상동 실험과 복제관찰 실험을 시행하여야 한다. 이러한 과정에서 적절한 표준화가 필

요하며 표준화과정은 S+통계소프트웨어를 통하여 생물학자도 구현할 수 있다. 그러나 자료에 따른 적절한 표준화법을 선정하는 것이 매우 중요하고 또한 실험의 원하는 결과를 효과적이고 정확하게 얻기 위하여, 실험의 설계부터 분석방법 등의 모든 과정에서 통계학자와 생물학자간의 긴밀한 협조와 상호이해가 필요하며, 생물학적 중요성과 통계적 유의성을 함께 고려하여 결론을 유도하는 지혜가 요구된다.

REFERENCES

1. Duyk GM. Sharper tools and simpler methods. Nat Genet 2002;32:465-468.
2. Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, Amico MD, Pestell RG, West M, Nevins J. Gene expression phenotypic models that predict the activity of oncogenic pathways. Nat Genet 2003;34:226-230.
3. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross G, Ladd C, Pohl U, Hartmann C, McLaughlin ME, Batchelor TT, Black PM, Deimling AV, Pomeroy SL, Golub TR, Louis DN. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res 2003;63:1602-1607.
4. Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton-Dutoit S, Wolf H, Orntoft TF. Identifying distinct classes of bladder carcinoma using microarrays. Nat Genet 2003;33:90-96.
5. Gibson G. Microarrays in ecology and evolution: a preview. Mol Ecol 2002;11:17-24.
6. Keshava N, Ong T. Gene expression patterns in human liver cells exposed to Tetrachloroethylene and its metabolite using microarray analysis (Paper #90). Environ Mol Mutagenesis Suppl 2003.
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-537.
8. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nat Genet 2000;403:503-511.
9. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Raffeld M, Yakhini Z, Ben-Dor A, Dougherty E, Kononen J, Bubendorf L, Fehrle W, Pittaluga S, Gruvberger S, Loman N, Johannsson O, Olsson H, Wilfond B, Sauter G, Kallioniemi OP, Borg A, Trent J. Gene-expression profiles in hereditary breast cancer. N Eng J Med 2001;344:539-548.
10. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR. Multicalss cancer diagnosis using tumor gene expression signatures. PNAS 2001;98:15149-15154.
11. Rosenwald A, Wright G, Chan Wc, Connors JM, Campo E, Fisher RI, Gascoyne RD, Konrad Muller-Hermelink H, Smeland EB, Staudt LM. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell Lymphoma. N Eng J Med 2002;346:1937-1947.
12. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrich M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. N Eng J Med 2002;347:1999-2009.
13. Speed TP. Gene expression data: Question, answers and statistics. Presentation at Bioinformatics 2001, Skovde, Sweden, March, 2001. http://www.stat.berkeley.edu/users/terry/zarrar/Html/talksindex.html
14. Rha SY, Kim TS, Jeong SJ, Ahn JB, Shim KY, Kong SJ, Lee HY, Yoo NC, Choi JH, Lim HY, Kim JH, Roh JK, Min JS, Kim BS. Effect on malignant phenotype of gastric cancer cell line after p53 gene transduction. J Korean Cancer Assoc 1998;30:508-520.
15. Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. J Com Graph Stat 2002;11(1):108-136.
16. Yang YH, Speed T. Design issues for cDNA microarray experiments. Nature Genet Rev 2002;3:579-588.
17. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology 2002;3(7):0036.1-0036.21.
18. Quackenbush J. Microarray data nomalization and transformation. Nat Genet 2002;32:496-501.
19. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contribution of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet 2001;29:389-395.
20. Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002;32:490-495.
21. Simon R, Radmacher MD, Dobbin K. Design of studies using DNA microarrays. Genetic Epidemiology 2002;23:21-36.