Comparison of Normalizations for cDNA Microarray Data

김 윤 희[1], 김 호[1], 박 웅 양[2], 서 진 영[3], 정 진 호[3]

Abstract

cDNA microarray experiments permit us to investigate the expression levels of thousands of genes simultaneously and to make it easy to compare gene expression from different populations. However, researchers are asked to be cautious in interpreting the results because of the unexpected sources of variation such as systematic errors from the microarrayer and the difference of cDNA dye intensity. And the scanner itself calculates both of mean and median of the signal and background pixels, so it follows a selection which raw data will be used in analysis. In this paper, we compare the results in each case of using mean and median from the raw data and normalization methods in reducing the systematic errors with arm's skin cells of old and young males. Using median is preferable to mean because the distribution of the test statistic (t-statistic) from the median is more close to normal distribution than that from mean. Scaled print tip normalization is better than global or lowess normalization due to the distribution of the test-statistic.

Keywords : cDNA microarray, print-tip, scale, lowess, normalization,

1. Introduction

수 천개 유전자들의 발현을 동시에 분석할 수 있는 cDNA microarray 실험을 하는 것은 특정 집단의 세포 또는 조직이 유전자 발현 수준에 따라 차이가 나는 것을 찾아 비교 가능하게 하였다. control에 해당하는 cDNA에는 green 형광 표지를 하고 관심이 있는 cell의 cDNA에는 red 형광 표지를 한 후, 발현정도를 표시하는 형광의 강도에 따라서 gene의 발현 정도를 측정하여 비교하게 되는 것이다. 현재, 이러한 장점 때문에 이 실험을 이용하여 얻은 자료로 많은 분석이 이루어지고 있다.

그러나, 실험과정 자체가 매우 정교하기 때문에 microarrayer나 scanner등의 실험 기계를 통해서 자료 값들을 얻는데, 이때 기계적 오차(systematic variation)가 생겨나게 된다. 예를 들면, microarray에 gene 조각들을 hybridization 시키는 경우, 각 print-tip 에 따라서 구획이 나뉘어지게 되는데, 그 구획에 따라 기계가 여러번 움직이면서 재조정되므로 실험조건에 있어서

1) Dept. of Biostatistics & Epidemiology, School of Public health, Seoul National University, E-mail : nina78@snu.ac.kr ; hokim@snu.ac.kr
2) Dept. of Biochemistry, Seoul National University College of Medicine, Email : wypark@snu.ac.kr
3) Dept. of Dermatology, Seoul National University College of Medicine, Email : jhchung@snu.ac.kr
1)-3) Seoul National University, 28 yungon-dong, chongno-gu, Seoul, 100-799, Korea

print-tip들간의 variation이 있을 수 있고, 제조 회사에 따른 형광시약 간의 차이와 같은 제조 회사라 하더라도 green과 red 형광시약 간의 농도 차이가 있을 수 있다. 실제로 red 형광시약을 green 형광시약 보다 좀 더 많이 넣어주어야 둘 간의 농도가 적정량의 조화를 이룬다고 보고되어 있다. (Yang, 2001)

이렇듯, 실험 외적 조건에서 나타나는 오차는 통계적 계산을 통해 보정해 주어야 하는데, 보정해주는 방법 중 하나가 raw data들의 log 값에 대한 normalization이다. 자료의 설명을 위하여 red 형광에서의 빛의 세기를 R, green 형광에서의 빛의 세기를 G라고 하고, $M=\log_2(R/G)$, $A=\{\log_2(GR)\}/2$라고 했을때, *normalization* 방법에는 전체 M의 값을 N(0,1)로 보내는 global(median) normalization, A의 값을 고려하는 lowess normalization, print-tip 구획을 고려하는 print-tip normalization 마지막으로 각 print-tip의 scale까지 보정하는 scaled print-tip normalization 등이 있다. 이들을 비교해서 gene expression level의 t값 차이가 확실하게 나타나도록 보조하는 좋은 normalization 방법을 찾는 것에 목적을 두고 있다.

또, scanner를 통해서 나오는 signal과 background의 값을 읽을 때 resolution이나 pixel에 관한 기술적인 문제로 읽혀지는 값들에 이미 variation이 포함되게 된다. 따라서, scanner에서는 microarray에서 intensity를 읽을 때 정사각형 모양의 pixel로 나누어 읽고 이 수치들의 mean과 median 값을 output으로 내보낸다. 이번 연구를 통해서 두 가지의 값 중 어떤 것을 이용하면 이후에 두 집단 간에 큰 차이를 보이는 gene의 t값을 구하여 비교할 때 더 정규분포에 가까워지는 지 알아보도록 한다.

2. Data

본 연구에서 사용한 시료는 70세 이상의 남성 2명과 20대 남성 3명을 대상으로 상완 피부 중 안쪽 부위 생검 조직을 control로 하고 햇빛에 노출되는 바깥 쪽 부위 생검조직을 target으로 정하였다. 여기서 얻어진 유전자 발현의 차이를 통해서 햇빛에 노출이 되는 바깥쪽과 안쪽 피부를 비교하여 광노화(photo aging) 효과를 알아보고자 하는 것이 이 연구의 목적이다. DNA repair, cell cycle, metabolism 등의 유전자들과 기능을 알지 못하는 EST에 해당하는 유전자들을 포함한 4608개 유전자를 이용하여 만든 cDNA microarray를 사용하였다. 총 5명에서 얻어진 시료들은 모두 microarray slide에서 왼쪽과 오른쪽에 반복하여 측정하였는데, 반복 측정한 것이어도 둘 간의 강한 연관성을 보이지 않으므로($0.4<r<0.65$, r=Pearson correlation coefficient) 다른 개체에서 온 것으로 간주하여 분석에 임하였다. 이는 연구 대상자가 70대에서 2명, 20대에서 3명으로 총 5명으로 너무 적은 sample의 수를 가지는 것을 보완해 주고 있기도 하다. gene들간의 연관성을 감안하여 70대와 20대, 두 집단의 비교시 permutation t-test를 사용하여 가장 큰 t값의 차이를 보이는 100개의 gene을 비교 검토해 광효과(photo aging)에 관여하는 gene을 찾아내는 것이 진행되고 있는 연구의 궁극적인 목표이다.

3. Method

cDNA chip microarray 실험자료가 많아지는 추세에 따라 gene data처럼 방대한 크기의 자료를 다룰 수 있고, 이를 용도 별로 분석할 수 있는 statistical software R program (from http://cran.r-project.org) 중 SMA package를 이용하였다. 전체적으로는 target group과 control group의 gene 비교이므로 t-test에 의거하게 되는데, 이때 gene들의 생물학적 상호 연관성을 감안하여 permutation t-test를 실행한다. j번째 gene의 normalized target 집단의 평균과 control 집단의 평균을 이용하여 t값을 계산하고, random permutation을 한 후 아래 (1)의 식을 이용하여 t-statistic을 계산한다.

$$t_j = \frac{\overline{x_{2j}} - \overline{x_{1j}}}{\sqrt{\frac{s^2_{1j}}{n_1} + \frac{s^2_{2j}}{n_2}}} \quad (j=1,2,\cdots,4608) \qquad (1)$$

여기서 얻어진 분포로 adjusted p-value를 이용하는 것이 두 집단에서 큰 차이를 보이는 gene의 판단기준이 된다. 이때 log ratio data에 normalization을 해 주게 되는데, 본 논문에서 고려된 4개의 normalization 일반식은 table 1과 같다.

Table 1. Equations of 4 different methods of Normalization

method	Equation
Global normalization(median)	$\log_2 R/G - c = \log_2 R/(kG)$, c= median or mean of the log-intensity ratios
Intensity dependent normalization (lowess)	$\log_2 R/G - c(A) = \log_2 R/(k(A)G)$, A=log average ratios
Print-tip normalization	$\log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G)$, i=print-tip
Scaled print tip normalization	adding scale factor(a_i) for the ith print-tip group

4. Results

4.1 Comparing Median and Mean of raw data

pixel을 이용해서 읽혀지는 signal과 background intensity 값에 대한 두 가지 통계량 median과 mean의 비교는 t statistics로 이루어진 분포가 얼마나 정규분포에 근사하는 지와 반복 실험의 correlation 값으로 판단한다. (Figure 1.1-1.4, Table 2)

Figure 1.1에서 Figure 1.4를 살펴보면 Normalization을 했을 경우, median을 사용한 것(Fig.1.4)이 mean을 사용한 것(Fig.1.2)보다 분산이 작은 것을 눈으로 확인 할 수 있다. 유효한 gene을 판단하는 기준이 되는 분포이므로 이후 통계적 계산을 위해서는 정규분포에 근사함을 보이는 분포를 선호해야한다. 그리고, Figure 1.1과 Figure 1.3의 Normalization을 하지 않았을 때의 분포를 보면 중심이 많이 치우쳐 있는 것을 볼 수 있다. 따라서, Normalization의 필요성을 보여주고 있다.

Table 2. Pearson Correlation Coefficient between replicated slides

	slide	Using Mean	Using Median
Before Normalization	old 1	0.58	0.63
	old 2	0.53	0.57
	young 1	0.48	0.69
	young 2	0.60	0.62
	young 3	0.54	0.63
After Normalization	old 1	0.42	0.65
	old 2	0.31	0.49
	young 1	0.35	0.60
	young 2	0.43	0.60
	young 3	0.36	0.64

※ old i : old sample ith slides(L/R) , young i: young sample ith slides(L/R)

Figure 1.1 Histogram and QQ plot of t-statistic using mean before normalization

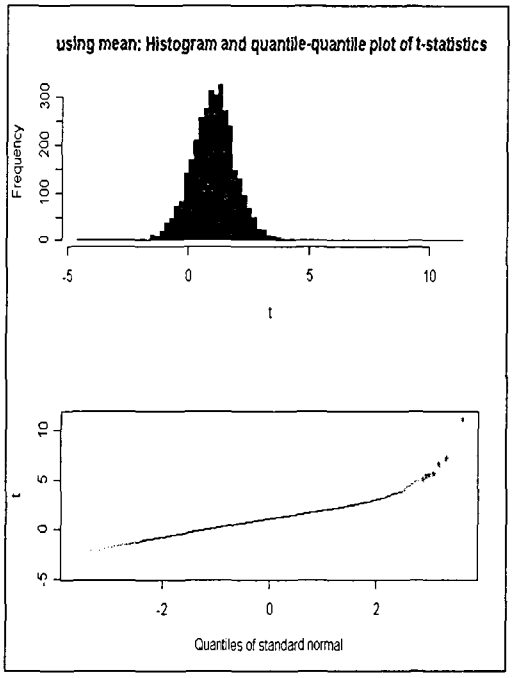

Figure 1.2 Histogram and QQ plot of t-statistic using mean after scaled print-tip normalization

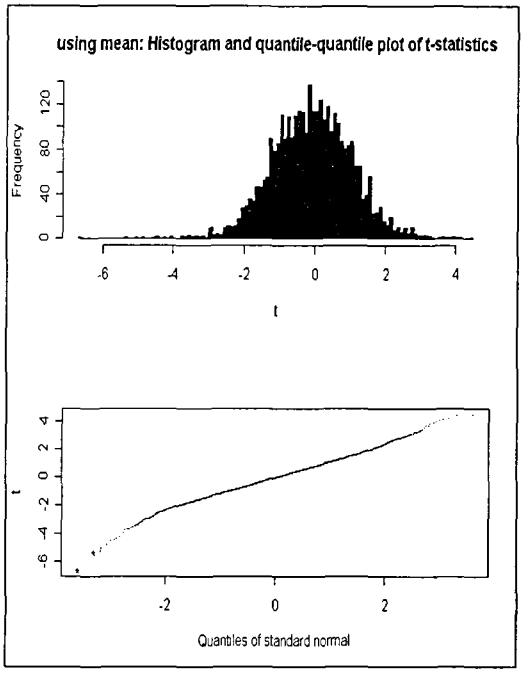

Figure 1.3 Histogram and QQ plot of t-statistic using median before normalization

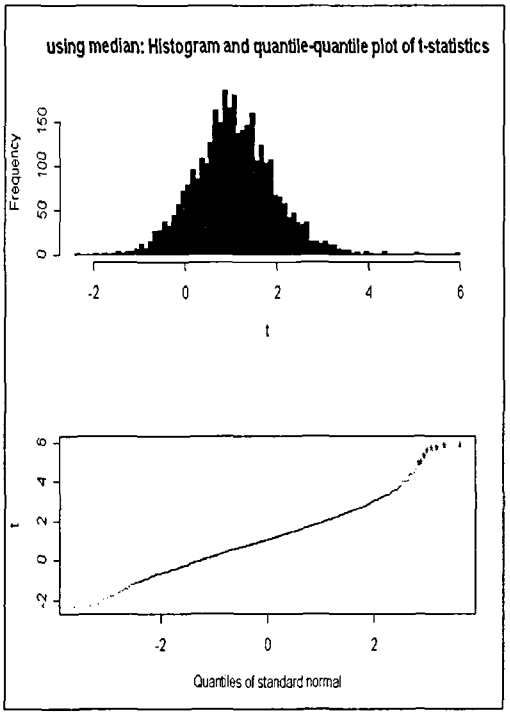

Figure 1.4 Histogram and QQ plot of t-statistic using median after scaled tip normalization

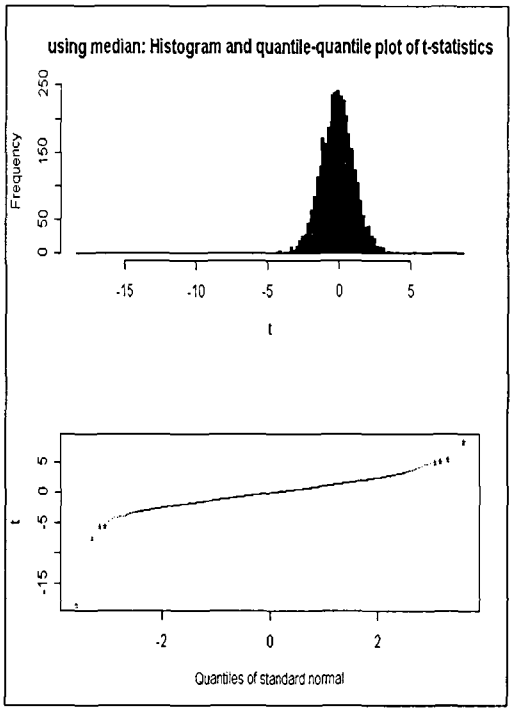

Table 2에서 나타내고 있는 것은 replication의 Pearson r 값들로 Normalization을 했을 경우와 안 했을 경우를 모두 보여 주고 있다. 이때 mean을 사용했을 때보다 median을 사용했을 때 r 값이 조금씩 증가 한 것을 볼 수 있다. 시간 차이 만 있을 뿐 같은 기계에 의한 반복실험이었으므로 한 개체에 대해서 반복 실험의 correlation이 높아야만 한다. 하지만 mean을 분석에 사용하였을 때는 0.5 정도로 낮은 수치를 보이다가 median을 사용하였을 때는 0.65정도로 증가하는 것을 볼 수 있다. 이 차이는 normalization을 했을 때에 더 크게 나타나고 있다. 이는 기계적 오차를 줄이는 것에 대해서는 median을 사용하는 것이 variation을 줄이는 데에 도움이 된다는 근거이다.

4.2 Normalization methods

5가지 normalization 방법들을 비교하기 위해서는 각 방법들에 따라 M값의 density plot (Figure 2)와 M vs A plot (Figure 3)들을 비교한다. Figure 2에서 보면, Normalization을 하지 않았을 때보다 Global normalization은 0쪽으로 위치이동이 되었고, Lowess normalization일 경우, M값 분포가 정규분포 모양에 근사한 것을 볼 수 있다. Print-tip normalization을 한 경우, variation이 줄어들어 정규분포의 꼬리가 짧아진다. Scaled print tip normalization 일 때는 전자의 경우보다 약간 더 variation이 작아 지는 것을 볼 수 있다. 마지막 그림은 위의 다섯 경우의 density plot을 비교를 위하여 모두 같이 그린 그림이다. 따라서, variation이 가장 작게 되는 scaled print-tip normalization이 가장 적합한 normalization 방법임을 알 수 있다.

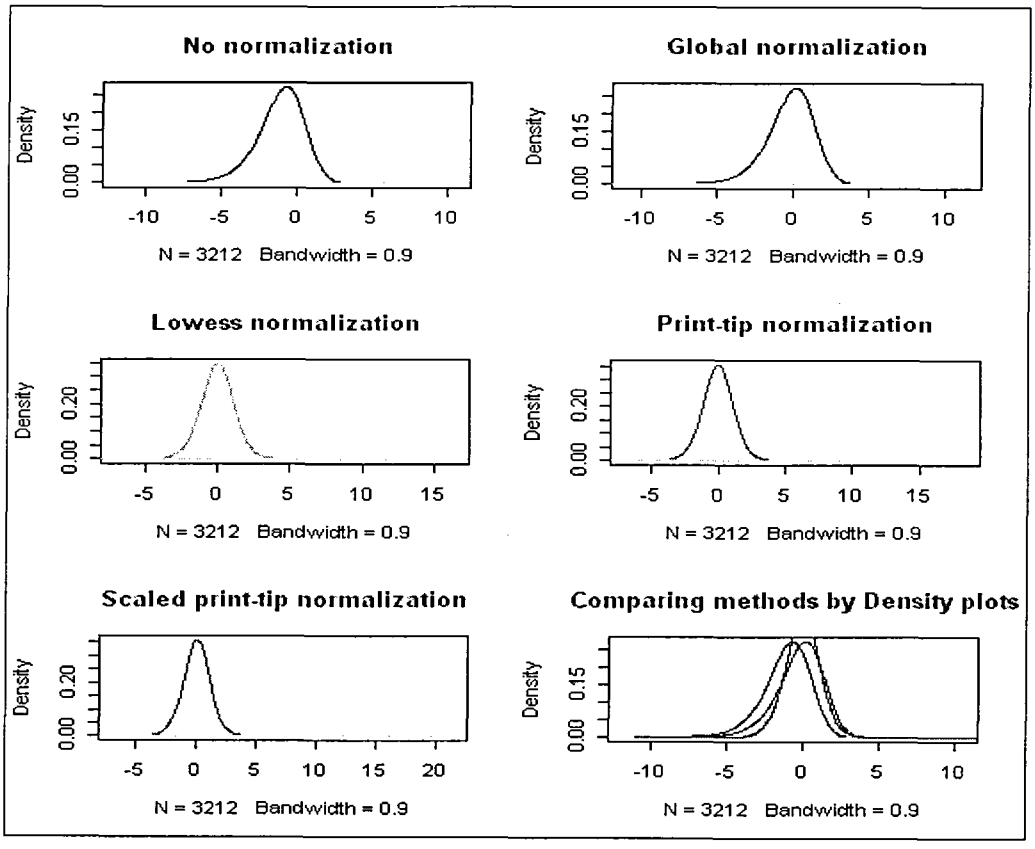

Figure 2. Density plots of different methods(4) & Combination plot : data from slide #4

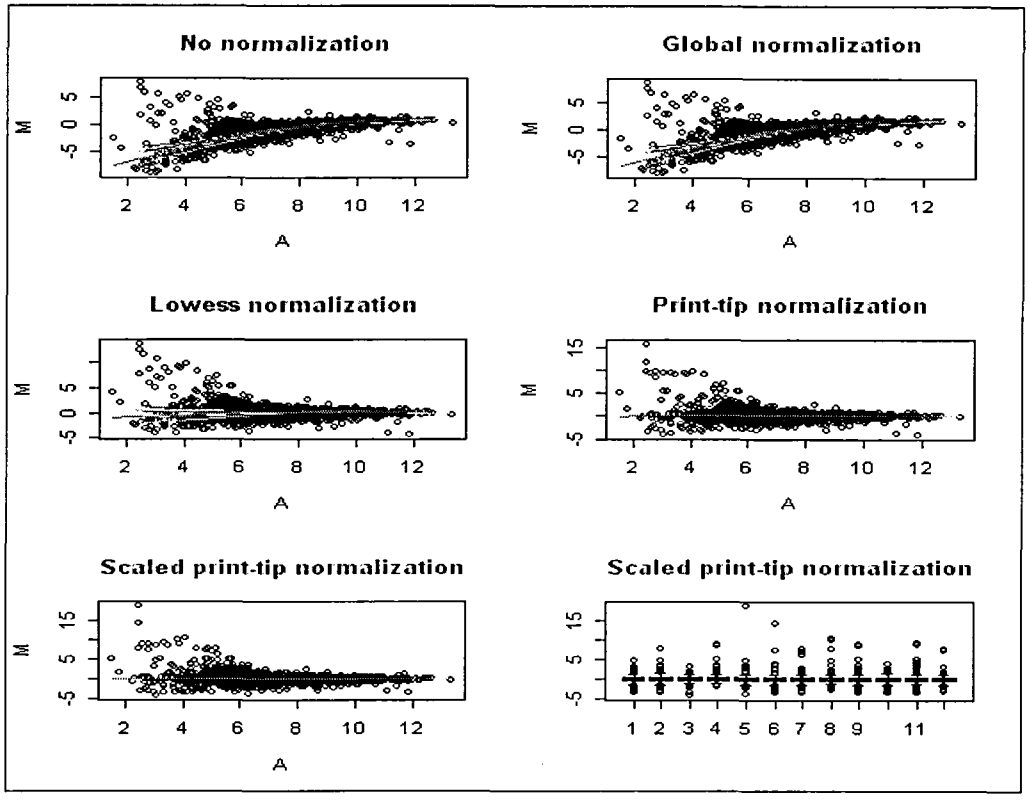

Figure 3. M vs A plot by method of Normalization & Scatter plot of Scaled print tip normalization
: data from slide #4 (old sample 2 right slide)

Figure 3의 M vs A plot 또한 normalization 방법을 비교하기 위한 그림인데, 이 실험에서 쓰인 slide의 print-tip 12개 각각의 lowess line을 한눈에 보여주고 있다. 기계적 오차를 줄이기 위한 normalization이 필요 없다면, 형광의 강도(intensity)는 같은 비율로 올라가야 한다. 즉, log G와 log R의 평균값인 A가 변하여도 항상 log G와 log R의 비율인 M값은 일정해야 한다. 따라서, M vs A plot에서 각 print tip의 lowess line들이 수평선을 가지게 되는 것이 가장 적합한 normalization 방법인 것이다. normalization을 하지 않았을 경우, 굽은 lowess line들이 print-tip normalization과 scaled print-tip normalization에서 곧은 직선으로 나타나는 것을 볼 수 있다. 마지막으로 scaled print-tip normalization을 했을 경우의 scatter plot은 각 print-tip 별 variation을 비교할 수 있는 그림이다. 각 12개의 print-tip 별로 고른 variation을 가지게 되는 것을 확인할 수 있다. Fig.2와 Fig.3을 통해서 scaled print-tip normalization이 효과적인 방법임을 알 수 있다.

5. Discussion

이상으로 cDNA microarray data들을 가지고 분석에 들어가기 전에 기계로 인하여 생겨난 오차를 최소화하기 위해 다루어야 할 과정을 살펴보았다. 우선 pixel들로 읽혀지는 값들의 mean과 median 값에 대한 비교에서는 median값이 반복에 대한 correlation도 높았고, 분석에 이용하는 t statistics 분포도 정규분포에 가까운 것으로 나왔다. intensity를 읽어들이는 과정에서

signal과 background에 대한 variation 범위가 넓기 때문에 outlier의 영향을 덜 받는 median의 사용이 타당한 결과가 나왔으리라 생각한다.

자료들의 normalization 또한 기계 오차를 통계적 계산으로 최소화하기 위한 방법으로 본 논문에서는 4가지를 제시하였다. log ratio 값인 M 값의 분포를 정규 분포로 근사시키기 위하여 위치이동과(global normalization) 분산의 최소화(lowess normalization)를 모두 포함한 scaled print-tip normalization이 가장 적합한 normalization 방법이었다. 각 slide내에서 한번의 기계 움직임으로 제어되는 print-tip에 의한 구획이 normalization에 큰 영향을 주고 있음을 알 수 있었다. 실험 환경이 동일하더라도 시간 차이까지 감안할 수는 없으므로 동시에 움직이는 print-tip을 고려 해주는 것이 올바른 normalization 방법으로 유의한 유전자를 밝히는 이후 분석에서 조금 더 확실한 결과를 보여 줄 수 있을 것이라고 본다.

자료의 수가 방대할수록, 반복 수가 적을수록 자료 정리가 요구되는 분석이 바로 cDNA microarray 실험에 관한 분석이다. 새로운 사실 발견에 급급한 나머지, 자료 내부에서 생겨난 오차를 간과하여 올바르지 않은 사실을 받아들이는 실수를 범하지 않도록 주의해야하겠다.

References

[1] Yang, Y. H., Dudoit, S., Luu, P., D. M., Peng, V., Ngai, J., and Speed, T. P.,(2001). *Normalization for cDNA microarray data,* SPIE BiOS 2001, San Jose, California.

[2] S. Dudoit, J. Fridlyand, and T. P. Speed (2000). *Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data*

[3] S. Dudoit, Y.H. Yang, M. J. Callow and T.P. Speed (2000). *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.*

[4] Ingrid Lonnstedt, Terry Speed(2001)., *Replicated microarray data*

[1]-[4] from Technical report, Department of Statistics, University of California at Berkeley http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html

[5] Andreas Krause, Melvin Olson(1997), *The Basic of S and S-Plus*, Springer, New york, USA