Genomics & Informatics Vol. 2(2) 92-98, June 2004

An Iterative Normalization Algorithm for cDNA Microarray Medical Data Analysiss

Yoonhee Kim[1], Woong-Yang Park[2] and Ho Kim[1]*

[1]Department of Biostatistics and Epidemiology, School of Public Health, and Institute of Health and Environment, Seoul National University, 28 Yungon-dong, Chongno-gu, Seoul, 100-799, Korea

[2]Department of Biochemistry Seoul National University College of Medicine,28 Yungon-dong, Chongno-gu, Seoul, 100-799, Korea

Abstract

A cDNA microarray experiment is one of the most useful high-throughput experiments in medical informatics for monitoring gene expression levels. Statistical analysis with a cDNA microarray medical data requires a normalization procedure to reduce the systematic errors that are impossible to control by the experimental conditions. Despite the variety of normalization methods, this paper suggests a more general and synthetic normalization algorithm with a control gene set based on previous studies of normalization. Iterative normalization method was used to select and include a new control gene set among the whole genes iteratively at every step of the normalization calculation initiated with the housekeeping genes. The objective of this iterative normalization was to maintain the pattern of the original data and to keep the gene expression levels stable. Spatial plots, M&A (ratio and average values of the intensity) plots and box plots showed a convergence to zero of the mean across all genes graphically after applying our iterative normalization. The practicability of the algorithm was demonstrated by applying our method to the data for the human photo aging study.

Keywords: cDNA microarray,normalization, iterative algorithm, biostatistics, medical data, control genes

Introduction

As a high-throughput technology, a cDNA microarray experiment that can simultaneously monitor the expression levels of thousands of genes has been widely used after the success of determining the human DNA sequences. However, there are limitations in that only few samples can be obtained *in vivo* and there is no golden rule for verifying the results and effects of the experiments. Therefore, reliable experiments and considerate analyses to reduce the errors in the results are needed. Regarding experiments, laboratory technicians can prevent intra and inter-laboratory errors by observing a precise protocol. However, other errors that cannot be controlled through physical treatments, such as errors from the mechanics or from experimental materials, demands the statistical adjustments during the analysis. In order to solve these problems, normalization is an essential part at the preliminary stages of an analysis. In Medical informatics, normalization implies the removal of errors and making a randomized distribution of the M (intensity ratio of expression levels between control and target genes) values. The final goal of normalization is to balance the different intensities of the two dyes in a slide caused from labeling or scanning sets. This should dispose of the errors and make a clear randomized data set to analyze reliably. Specifically, after a good normalization, the intensity ratio value (M)might show a close mean of M near zero at each average intensity level (A). If there is a systematic error, the M value across the A value may have a curvature on the M&A plot. This is because the systematic errors make a certain part of the data have low or high values particularly. Normalization can make all M values even and randomly along by A values. Therefore, an M&A plot should have an almost straight band of M values.

During the procedure of normalization, there are several ways for correcting errors, for example, regression fitting analysis, and ANOVA as a calculation method. Additionally choosing the control gene group for normalization is also a controversial issue. Slide-wise, pin-wise and average pin-wise normalization procedures all focus on each spot on the scatter plot over the slides (Schuchhardt *et al.*, 2000) .In addition, the total intensity normalization method assumes that the total integrated intensity computed for all the samples is the same and simply subtracts the mean. Likewise, Yang *et al.*(2002) suggested a global normalization method with the whole gene set. Normalization used regression techniques

*Corresponding author: E-mail hokim@snu.ac.kr, Tel +82-2-740-8874, Fax +82-2-745-9104
Accepted 16 April 2004

expect an adjustment of the slope to a straight line in a scatter plot of a Cy3 and Cy5 intensity (Hedenfalk *et al.*, 2001). LOWESS (Locally Weighted Scatterplot Smoothing) is a popular method for making a local adjustment (Cleveland and Devlin, 1988). Chen *et al.* (1997) considered control gene sets at the ratio statistics of normalization. Yang *et al.* (2002) developed normalization methods under the consideration in a structure of arrays and reported a print-tip group normalization method. Zien *et al.* (2001) disproved the global normalization and used a housekeeping gene approach. They instead proposed a pair-wise scaling method for centralizing all the genes over the slides. In contrast, Wang *et al.* (2002) reported the necessity of control gene set and presented an iterative method using control genes. With this theoretical background, the aim of this paper was to compensate for some of the drawbacks of previous studies and propose a novel normalization algorithm in a universal and synthetic way.

Materials and Methods

System

Biologically, housekeeping genes have still been thought as reliable biomarkers i.e. they are available as a consistent control gene set in a cDNA chip. Therefore, the normalization method used in this paper also used housekeeping genes with all the genes as a control gene set in the calculation procedure. Furthermore, updating a new control gene set among all the genes on the same slide iteratively at each step is an essential point of this study.

Some normalization methods are able to force the data to transform their values at once (Wang *et al.*, 2002). In that point of view, the LOWESS method with newly selected control genes at each iterative step can be a more moderate, robust and safe procedure. The iterative normalization method emphasizes the removal of systematic errors by maintaining the original pattern of the raw data. The systems of iterative normalization method follow two points: First, in the normalization procedure, the housekeeping genes as well as all the genes on a slide are used as a control gene set. Second, by the iterative calculations based on the LOWESS method, normalization is naturally accomplished.

Materials and methods

On a cDNA chip, there are two identical slides for an internal comparison. Each slide consists of twelve arrays accounting for the print-tips (Yang *et al.*, 2002) and a total 4608 human cDNA probes related to the DNA repair, cell cycle, metabolism and unknown ESTs, etc.. This means there are 12 x 1 arrays on one slide and 12 x 32 spots (sub-grids) on a single array. At each array, the last row consists of eight repeated sets of control genes; actin, lambda 564, and tublin. These genes are the housekeeping genes including negative and positive controls. Data analyzed in this paper were designed for a case study of the photo-aging effect of the skin. The reference sample was extracted from an inner fore arm skin biopsy and the target mRNAs from the outer fore arm skin biopsy were exposed to more sunlight. By comparing the reference and target samples, the effect of sunlight can be determined. Once the biological experiment was complete, a scanner reads the fluorescent intensity on a chip. After reading, it records the mean, median and mode of the signal and background pixel intensities automatically. Among these values, the median is a good estimator of representing the intensities considered to have a pixel property. Therefore, the median value of the signal and background intensities is managed as the raw data in the calculation. The SMA packages in R software (*http://cran.r-project.org*) based on S-plus were used in the analysis.

Algorithm of iterative normalization

This section shows the new Iterative normalization algorithm with the following steps.

1. Set the initial control gene set with the immobilized known housekeeping genes.
2. Apply the LOWESS to the control gene set and obtain the $\hat{M}^{i}_{j(C)}$ estimate values.
   (i = iterative number;1,2,3,$\cdots$, j(c) = gene number in control gene set; 1,2,$\cdots$,k)
3. Approximate the $\hat{M}^{i}_{j}$ (j=gene number;1,2,3,$\cdots$,N) value across the A values of all genes along the expanded LOWESS estimates ($\hat{M}^{i}_{j(C)}$).
4. Make a new M^{i}_{j} value by the subtracting $\hat{M}^{i}_{j}$ value from the previous M^{i-1}_{j} value.
5. Extract the new control gene set within ±0.05 boundaries of the M^{i}_{j} value approximately 8% ($=\frac{1}{12}$) of the whole genes.
6. Repeat the loop through process 2 to 5 iteratively until S_i converges to 0.

$$S_i = \sum_{j=1}^{N} \left| M^{i}_{j} - M^{i-1}_{j} \right|$$

(i = iterative number; 1,2,3,$\cdots$ j= gene number;1,2,3,$\cdots$.N)

The main idea of iterative normalization is to combine the LOWESS method and the iterative calculation by considering the control gene set. This means that the

iterative normalization method focuses on the best use of all the genes on the same slide as a control set including the housekeeping genes initially in the process 1 and applying the iterative LOWESS methods to processes 2 through 5. Accordingly, it can prevent data distortion that can be caused by putting data of other sources compulsorily in the adjustment. Actually, the key point is to renew the control gene set from a homogenous environment, such as the same experimental time and same laboratory worker, from being on the same slide. At process 5, the reason for selecting stable genes staying near zero area at approximately 8% ($=\frac{1}{12}$) of all genes is to reflect the size of the initial control gene set size. As mentioned before, 384 housekeeping genes among 4608 genes were immobilized on a cDNA chip. In order to extract the stable control genes, approximately 8% of all genes, the genes were selected within the ±0.05 boundaries of 0 in the M_j^i values, which is a size that satisfies approximately 8% of the whole genes empirically. If the difference in M_j^i and M_j^{i-1} converges to 0, it means that the positions of the genes are stable and almost fixed. Therefore, normalization keeps processing until the index S_i has the same value over the iterations.

Results

Many statisticians have proposed new methods using several assumptions. Indeed, there is no robust model to evaluate the various normalization methods. Some people prefer numerical and logical formulations. However, graphical evaluating method has attracted a great deal of attention. SMA packages of R provide the function to display some of the intensity of the spots on the array, such as a scanning image by the shades of gray or colors. These spots represent the M values within an absolute critical value defined with quality information such as the spot size or shape. Using this spatial plot, any spatial effect in the data can be examined visually. Usually, the scatter plot of M and A (M&A plot) with the LOWESS line of data is preferred because it enables up the pattern of whole genes to be checked at a glance. Along the average of the intensities, A, on the X-axis, the intensity ratio, M, is expected to have the same mean value, zero. In order to illustrate this, M cannot be affected by the A values even though the target and reference sample intensities (R and G) have low values inducing low A values. This means that M is always meant to spread concentrically and symmetrically across all A values. Besides, a box plot separated by print-tips can be used to explore the print-tip effect. It shows comparable box plots of each print-tip in a row. After seeing this plot, the mean and variance of the M values confirms the higher specificity than the M &A plot. The plots as above are ready to evaluate the normalization method. Using a comparison of each plot before and after normalization, the effect of normalization will be verified approximately. In this paper, all three plots were used as results for evaluating the iterative normalization method from one slide of a data set in the cDNA microarray experiment.

Spatial plot

Iterative normalization has performed for 25 times iterations and was stopped when S_{25}=0.75≈0 was sufficiently satisfied. (Table 1) Through each iterative step, the S values were reducing because the changes in the M values decreased gradually. Coming to the end of iterations, there are fewer changes in the M values and those values become almost fixed. Figure 1 shows that the S value from every iterative stage makes a declining line, which ensures that the S value converges to zero.

Table 1. S values at each iterative step

Iteration Number=i	Si	Iteration Number=i	Si
1	9.90	14	1.63
2	5.85	15	1.82
3	3.76	16	1.66
4	4.50	17	1.19
5	4.49	18	1.29
6	3.49	19	2.04
7	2.91	20	1.60
8	2.55	21	1.21
9	1.98	22	0.92
10	1.30	23	0.81
11	2.10	24	0.76
12	1.87	25	0.74
13	1.98		

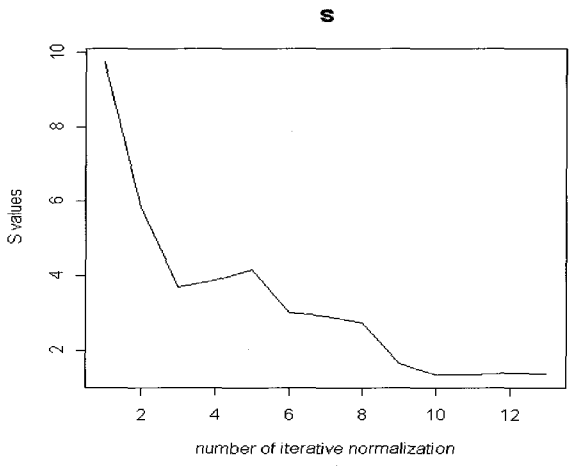

Fig. 1. Plot of S values at every iterative stage

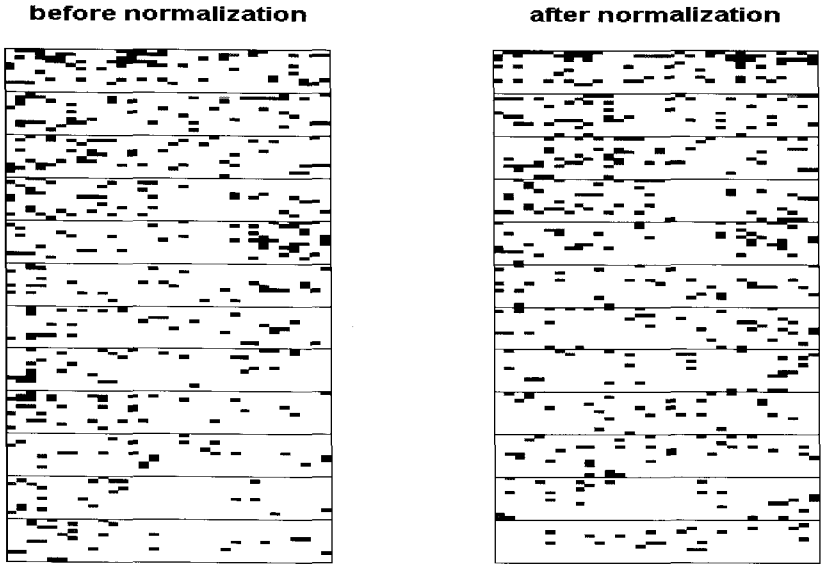

Fig. 2. Spatial plots before(left panel) and after(right panel) iterative normalization

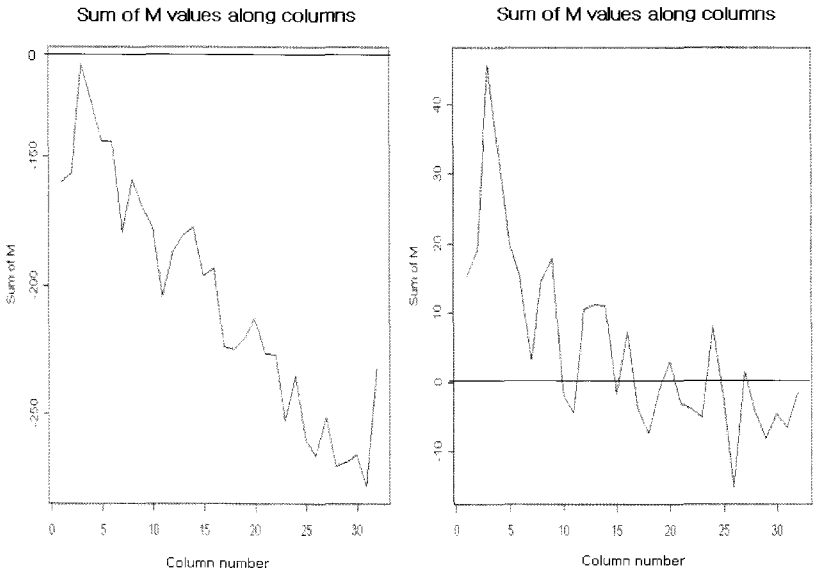

Fig. 3. Histograms along the columns before and after normalization

A spatial plot (Fig.2) prior to normalization illustrates the spatial problem caused by the mechanical procedure of the experiment. Red spots can be seen crowded around the left part on the slide. However, after iterative normalization, the red spots are located evenly all over the slide. In addition, the histograms (Fig.3) of the sum of the M values by column (32 sub-columns) on the spatial plot show the bias of the absolute M values within 10%. The left panel of Fig. 3, which shows a histogram prior to the iterative normalization, shows that the sum of the M values declines steeply under the zero line along the first column through the last one as seen in the pattern on the left plot in Figure 2. However, after normalizing, the sum of the M values by columns is up regulated near zero and appeared to have an equally wide band balanced near the zero reference line except for the left five columns.

A summary of the M values shows the improvement in the normalized distribution indicating a near zero value in both the mean (=0.04) and median (=0.1391) after the iterative normalization. (Table 2)

Table 2. The summary of the M values on the slide

Normalization	Before	After
Median	-1.334	0.1391
Mean	-1.374	0.0405

M & A plot

On the M&A plot before normalization (Fig. 4, first row), there is an abnormal curvature near the region 8 to 14 of the A values. Compared to the other average intensity, that region has slightly down regulated M values. However, the iterative normalization removes the errors, and the data is finally transformed into an intensity ratio that is distributed similarly over all the A values. In order to compare the results with the other normalization method, the third and fourth rows of Figure 4 show the M&A plots from the normalization with the housekeeping genes only (upper) and the global normalization (lower) (Yang *et al.*, 2002). The M&A plot of the global normalization still has the curvature that suggests that there should be more considerations in normalization not just subtracting the total mean value from the M values. In addition, the plot of the normalization with the housekeeping genes appears like the plot after the iterative normalization except that the former has a slightly wider variance than the latter. Using these results, we can be sure that the iterative normalization method is the best fit to this data set.

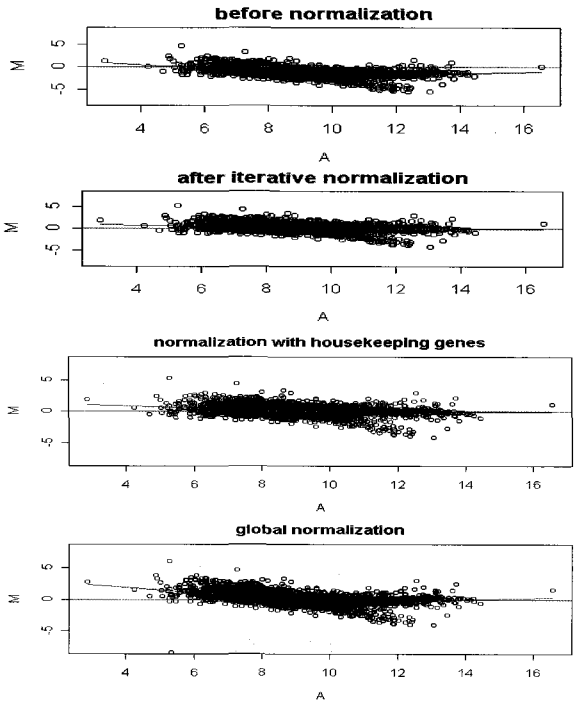

Fig. 4. M & A plots for several normalization methods

Box plot

The box plots separated by the 12 print-tip arrays may illustrate the print-tip effect. Prior to normalization, the median values of all the box plots are below the zero line of the M values. (Fig. 5) Since there is no unique high or low box plot, it can be assumed that the distinctive print-tip effect does not exist in this data. Meanwhile, it remains the problem that all values are down from the reference line (M=0). This situation may make it difficult to compare the M values with the other slides. After iterative normalization, median values of all the box plots are arranged approximately in a reference line. The data are adjusted by lifting up their M values in order for the original pattern to be totally maintained.

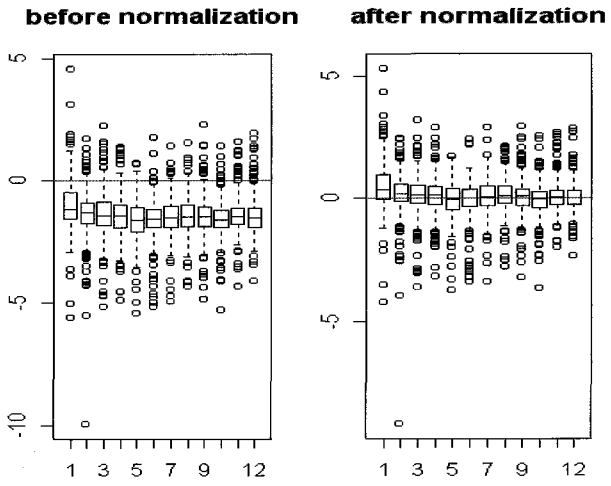

Fig. 5. The box plot along 12 print-tips

Discussion

Three kinds of plots - spatial plots, M&A plots and box plots showed satisfactory results after applying the iterative normalization method. A comparison spatial plots (Fig. 2) before and after normalization confirms the spatial effect arising from a certain mechanical problem. The spatial plot shows that the largest absolute M values above 10% are biased to the left part on a slide. At a guess, when the arrayer fixes the probes in that part of the slide, the intensity of the automated robot might have been different from the other part. In addition, the histograms (Fig.3) confirm the particular problem on the slideby the numerical values. This result showed the existence of a systematic error in prediction. After the iterative normalization method, the systematic error is relieved and the spatial plot displays balanced spot locations, and sum of M values are fluctuated and

remained almost zero relative to the ones prior to normalization. Therefore, iterative normalization removes the spatial effect from the systematic error.

The M&A plots with the LOWESS line (Fig.4) can grasp the whole pattern of the genes. After normalization, the curvature of the LOWESS line straightens to that of the reference line. For the raw data, the part of the curvature is located at higher A values. It is the opposite case of what is normally encountered. A common view of the intensities was that the curvature tends to occur at lower A values (Yang *et al.*, 2002). This situation informs us that cDNA microarray analysis requires careful scrutiny before applying the normalization method. In other words, the study of a more general normalization method without specific assumptions is more important. The iterative normalization method makes fewer assumptions than previous studies and worked its function successfully. On the M&A plot after normalization with the housekeeping genes only, the M values are somewhat similar to those of the Lowess line of M&A plot after iterative normalization but the density of the M values over the reference line is lower than that one. After a global normalization simply using the whole genes to normalize, the M&A plot has a bent Lowess line. This means that the unconditional control genes cannot perform any function to adjust for the errors. In conclusion, the iterative normalization algorithm has a more competitive result than the previous normalization method. Boxplots are good for illustrating the characteristics of the data such as the median, variance and outliers. Box plots, which were obtained here, reveal detailed descriptions about the print-tip effect. The whole raw data were down regulated from the reference line and this problem was adjusted after the iterative normalization. The box plot before normalization suggests that there were no abnormalities in the print-tips physically. After normalization, most of the median values approached those of the reference line (M=0). Therefore, the comparison of M values between print-tips is ready. Finally, the list of the control genes set of this study is available for assessing the biological implication. As the initial control gene set, 384 genes are repeatedly immobilized housekeeping genes. The first new control gene set has 406 genes including some housekeeping genes. After an Iterative normalization 25 times, 36 genes were omitted from the first control gene set and added to the other 28 genes. The results in this paper demonstrate that the iterative normalization method is suitable for normalizing and adjusting the raw data with systematic errors. The systematic errors were reduced and the normalization of the raw data can be accomplished through the iterative steps.

Conclusion

The Iterative Normalization Method uses both the homogeneous control genes as well as the housekeeping genes to normalize and not damage the whole pattern of raw data. To be sure that this method is suitable, more applications to other cDNA microarray data from other laboratories are recommended. Despite the same basis of the Microarray experimental protocol, there would be slight differences according to the particular laboratory such as arrayers. Nowadays, there are many web sites of groups involved in the DNA microarray experiments on the Internet that share data. Therefore, more applications using the data from these laboratories will be a good exercise to validate and evaluate the new normalization method. The use of different arrayers and different housekeeping genes is a cautious part to apply this method. Originally, the purpose of the cDNA microarray experiment was to identify the most significant genes of the target samples compared to the reference samples. Once the raw data is manipulated by normalization, reliable conclusions can be drawn to predict the significant genes that have a distinctive gene expression level by following the analysis e.g. SOM or clustering. Besides, if the cDNA microarray experiment has many samples, the normalization procedure will for the essential part for comparing the precision and validity of the samples. That is the absolute reason why its analysis requires a normalization of the raw data. There are some limitations in the iterative normalization method; it needs to iterate the loop in easy way and to find optimal conditions to stop iterations in programming.

Acknowledgements

This work was supported by grant No. R01-2002-000-00554-0 (2003) from the Basic Research Program of the Korea Science & Engineering Foundation.

References

Chen, Y., Dougherty, E. R., and Bittner, M. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed.Opt.* 2,364-374.

Cleveland, W. S. and Devlin, S .G. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* 83, 596-610.

David, J., Duggan, M. B, Yidong, C., Paul, M., and Jeffrey, M. T. (1999). Expression profiling using cDNA microarrays. *Nature genetics supplement.* 21, 10-14.

Dudoit, S., Fridlyand, J., and Speed., T. P. (2000). Comparison of Discrimination Methods for the Classification of

Tumors Using Gene Expression Data. Dept. of Statistics, University of California at Berkeley, Technical reports.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed., T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Dept. of Statistics, University of California at Berkeley, Technical reports.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M .L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439), 531-7.

Hedenfalk, L. *et al.* (2001). Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* 344, 539-548.

Kepler, T. B., Crosby, L., and Morgan, K. T. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol.* 3(7), RESEARCH0037.

Kim, J. H., Shin, D. M., and Lee, Y. S. (2002). Effect of local background intensities in thenormalization of cDNA microarray data with a skewed expression profiles. *Exp Mol Med.* 34, 224-232.

Michael, B., Eisen, P., and Brown, O. (2000). NA arrays for analysis of Gene Expression. Stanford University School of Medicine, Stanford, CA, Technical reports.

Quackenbush, J. (2001). Computational Analysis of microarray data. *Nature* 418-427.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* 28(10)E47.

Tran, P. H., Peiffer, D. A., Shin, Y., Meek, L. M., Brody, J. P., and Cho, K. W. (2002). Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res.* 30, e54.

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., and Wong, W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* 29, 2549-57.

Tsodikov, A., Szabo, A., and Jones, D. (2002). Adjustments and measures of differential expression for microarray data. *Bioinformatics* 18, 251-60.

Wang, Y., Lu, J., Lee, R., Gu, Z., and Clarke, R. (2002). Iterative normalization of cDNA microarray data. *IEEE Trans Inf. Technol. Biomed.* 6, 29-37.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng ,V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30, e15.

Zien, A., Aigner, T., Zimmer, R., and Lengauer, T. (2001). Centralization: a new method for the normalization of gene expression data. *Bioinformatics* 17(Suppl. 1), S323-31.