Genomics & Informatics Vol. 5(3) 118-123, September 2007

Effect of Normalization on Detection of Differentially-Expressed Genes with Moderate Effects

Seoae Cho[1], Eunjee Lee[1], Youngchul Kim[2] and Taesung Park[2]*

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-921, Korea, [2]Department of Statistics, Seoul National University, Seoul 151-921, Korea

Abstract

The current existing literature offers little guidance on how to decide which method to use to analyze one-channel microarray measurements when dealing with large, grouped samples. Most previous methods have focused on two-channel data;therefore they can not be easily applied to one-channel microarray data. Thus, a more reliable method is required to determine an appropriate combination of individual basic processing steps for a given dataset in order to improve the validity of one-channel expression data analysis. We address key issues in evaluating the effectiveness of basic statistical processing steps of microarray data that can affect the final outcome of gene expression analysis without focusingon the intrinsic data underlying biological interpretation.

Keywords: Normalization, CFS disease, Microarray, ANOVA

Introduction

Chronic fatigue syndrome (CFS) is known to be a complex disease related with several genes. In order to identify differentially-expressed genes, a large gene expression dataset was obtained using a one-channel microarray experiment conducted on 173 patients. The patients were classified into five groups of CFS, and 20,160 genes were represented. Unlike the usual microarray experiment, this study contained a very large number of slides. Unfortunately, however, our preliminary analysis yielded few significant genes differentially expressed among the five groups. This result was quite surprising, because 173 patients would be regarded as quite a large sample size in a typicalmicroarray experiment. Thus, we expect that the significant gene effect is quite moderate. Recently, microarray data interpretation has been mainly focused on the comparisons between the high density oligonucleotide-based chip and two-channel cDNA microarrays (Bolstad *et al.*, 2003; Edwards 2003; Futschik *et al.*, 2004; Cui *et al.*, 2003; Smyth *et al.*, 2003). Current literature offers little guidance on how to decide which method to use or how to compare different methods to obtain final results. These effects are most problematic, especially for one-channel microarray measurements when dealing with large, grouped samples (Edwards, 2003). In this paper, we focus on the effect of data processing on the interpretation of gene expression data for a one-channel microarray experiment when the gene effect is not large. We focus on evaluating the effect of normalization methods in identifying differentially-expressed genes.

Materials and Methods

Chronic fatigue syndrome (CFS) dataset

We analyzed a large gene expression dataset obtained from a one-channel oligonucleotide experiment conducted on 173 patients who were classified into five groups of chronic fatigue syndrome, and 20160 genes were represented. The data in this study describe CFS that has no diagnostic clinical signs. It is unclear if CFS represents a single illness (whistler *et al.*, 2005). The samples were classified into five groups based on information from clinical consensus: 36 patients in the non-fatigued (NF) group, 46 patients in the chronic fatigue syndrome (CFS) group, 47 patients in the group labeled chronically fatigued but without CFS because of an insufficient number of symptoms (ISF), 20 patients in the chronically fatigued but with ISF and a major depressive disorder with melancholic features (ISF-MDDm) group, and 19 patients with CFS with a major depressive disorder with melancholic features (CFS-MDDm). The NF group was defined as the control group, whereas the other four groups were defined as case groups based on various definitions of CFS. The final goal of the experiment was to identify differentially-expressed for the group pairs: control (NF) vs. case (all 4 groups);control (NF) vs case (CFS);control (NF) vs case (CFS-MDDm);control (NF) vs case (ISF);control (NF) vs case (ISF-MDDm);control (NF) vs case (ISF-MDDm+CFS-MDDm);and control (NF) vs. case (ISF+CFS). Findings from our study will suggest future studies needed to identify the underlying etiology of

*Corresponding author: E-mail tspark@stats.snu.ac.kr
Tel +82-2-880-8924, Fax +82-2-883-6114
Accepted 10 July 2007

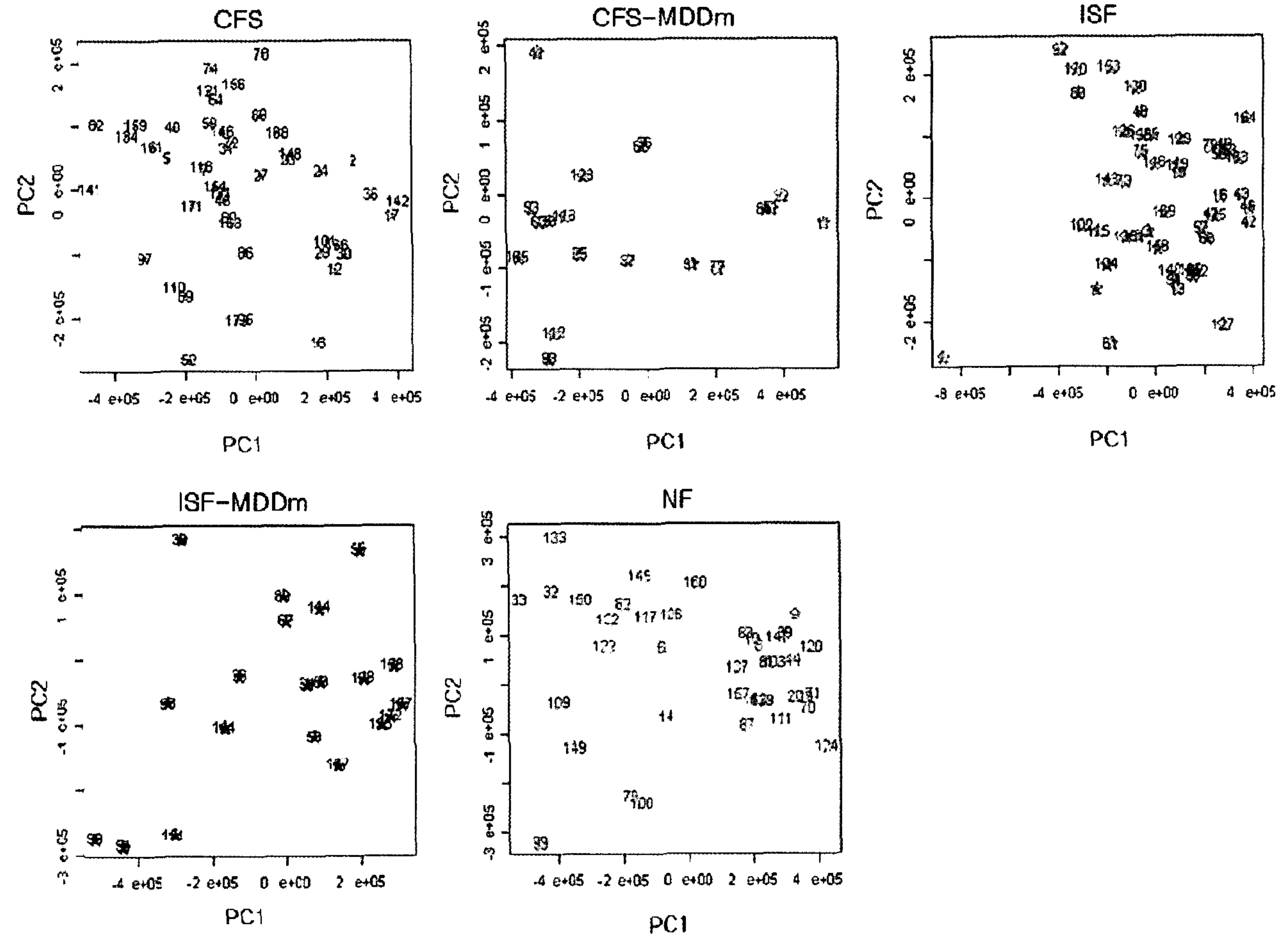


Fig. 1. Principal component analysis for each group. X-axis represent the scores of first principal component and Y-axis is the one of second principal component In this figure, outlying points indicate the outlier slides. Each plot is the PC plot of CFS, CFS-MDDM, ISF, ISF-MDDM, and NF groups.

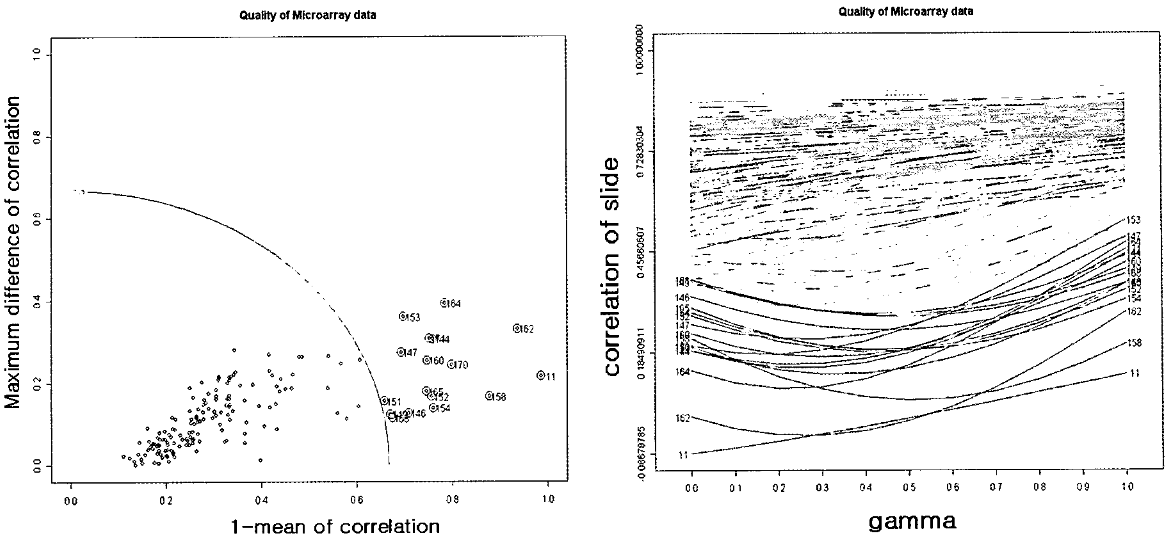
chronic fatigue syndrome;the dataset used for this study is not intended to focus on the intrinsic data underlying biological interpretation.

Quality Control Analysis and Outlier Detection

The first processing step is outlier detection, and we performed several methods to examine outlying slides. Fig. 1 is the result of the principal component analysis for the five groups, which identifies distinct outlying slides visually. We observed that samples 11 from the CFS-MDD group and 1 from the ISF group were separated from the dense group and could be classified as outlying slides. In addition, we examined slides with unusual expression patterns or large variability through diagnostic plots (Park *et al.*, 2005). These outlying slides tend to have large impacts on analyses such as the identification of differentially-expressed genes. Therefore, we applied graphical methods to detect outlying sample slides. With this measure of quality control, we were able to compare variability among slides of samples and minimize the amount of errors made in statistical data preparation of one-channel microarray data. Fig. 2 is the diagnostic plot for detecting outlying slides. The plot shows that slides 11, 162, and 158 seemed to have quite different patterns from those of other slides. Slide 11 was also identified by principal component analysis, showing a very clearly distinctive pattern from those of other slides. In order to evaluate the effect of the outlier slides, we performed leave-one out analysis. Finally, we used the dataset, removing these three outlier slides (slides 11, 162, and 158), for further analysis.

Normalization

The plot in Fig. 3 is the boxplot of the original intensity data from five individual sample groups before normalization. We



(a) Quality diagnostic plot

(b) Pattern of correlation over varying gamma values

Fig. 2. Quality control plot for detecting outlying sample slides. (a) shows that slides 11, 162, and 158 seem to have quite different patterns from those of other slides. Slide 11 was also identified by principal component analysis to have a very clearly distinctive pattern from those of other slides. We used the dataset after removing these outlier samples (slide 11,162,158) for further analysis.

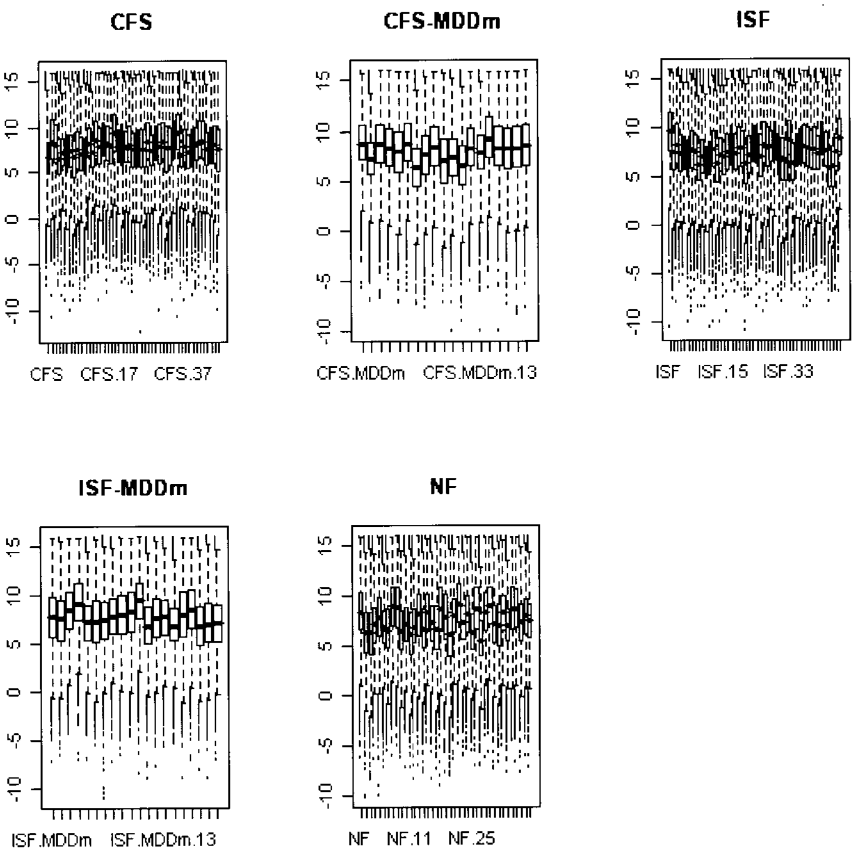
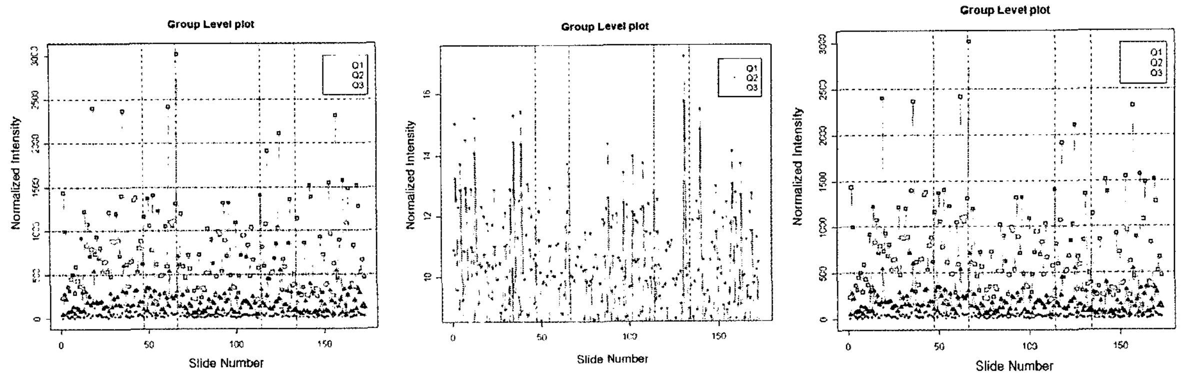


Fig. 3. Boxplot for raw data of each group and group-level plot (before normalization)



(a) Raw data (b) LOWESS normalization (c) Quantile normalization

Fig. 4. Group-level plot showing the effect of normalization. (a) shows the pattern of the 25th, 50th, and 75th quantiles of raw data from 173 slides. (b) is the pattern of the 25th, 50th, and 75th quantiles of LOWESS-normalized data. (c) is the pattern of the 25th, 50th, and 75th quantiles of quantile-normalized data.

performed the normalization process for each of the five individual sample groups as examined in Fig. 3. The first plot in Fig. 4 is the group-level plot for these raw data, showing the 25th, 50th, and 75th percentile for each sample of the five groups. The x-axis represents the slide, and the y-axis represents quartiles of the actual intensities. This group-level plot allows us to investigate group-specific variations within the data. The group-level plot clearly showed different levels of expression data for the different sample groups. The next processing step was the normalization procedure performed by Quantile, LOWESS, and IQR scale normalization methods (Yang *et al.*, 2002; Irizarry *et al* 2003; Li *et al.*, 2003).

Quantile normalization

The goal of the quantile method is to make the distribution of probe intensities for each array in a set of arrays the same.

Let $q_k = (q_{k1}, \ldots, q_{kn})$ for k = 1, ...,p be the vector of the kth quantiles for all n arrays. $q_k = (q_{k1}, \ldots, q_{kn})$and $d = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$let be the unit diagonal. To transform the quantiles so that they all lie along the diagonal, consider the projection of q onto d:

$$\mathrm{Proj}_d q_k = (\frac{1}{n}\sum_{j=1}^{n} q_{kj}, \cdots, \frac{1}{n}\sum_{j=1}^{n} q_{kj})$$

This implies that we can give each array the same distribution by taking the mean quantile and substitute it as the value of the data item in the original dataset. This motivates the following algorithm for normalizing a set of data vectors by giving them the same distribution:

1. Given *n* arrays of length p, form *Y* of dimension *p* × *n* where each array is a column.
2. Sort each column of *X* to give Y_{sort}.
3. Take the means across rows of Y_{sort} and assign this mean to each element in the row to get Y'_{sort}
4. Get $Y_{\mathrm{normalized}}$ by rearranging each column of Y'_{sort} to have the same ordering as original *Y*.

LOWESS normalization

Another approach is LOWESS normalization based on an *M* versus *A* plot, where *M* is the difference in log expression values and *A* is the average of the log expression values. To normalize two arrays with one-channel intensity, it is straightforward to adapt the approach proposed by Yang *et al.* (2001) for correcting dye bias in two-channel data, as follows. Let the log intensities from the two arrays be $Y^1 = (y_1^1, y_2^1, \ldots, y_P^1)$, $Y^2 = (y_1^2, y_2^2, \cdots, y_P^2)$. Consider the plot of $M = Y^1 - Y^2$ against $A = Y^1 + Y^2$, which corresponds to a clockwise rotation of the (Y^1, Y^2) plot by 45 degrees followed by rescaling and fit a locally weighted smooth regression (loess) $f(A)$ to these (M, A) data. The adjustment consists of replacing Y^1 by $\hat{Y}^1 = Y^1 - f(A)/2$ and Y^2 by $\hat{Y}^2 = Y^2 - f(A)/2$. Bolstad *et al.* (2003) extended this method to a series of k arrays, with data in the form $Y^1, Y^2, \ldots, Y^k$. However, rather than being applied to two-color channels on the same array, it was applied to sample intensities from two arrays at a time. Because this method works in a pairwise manner, it is somewhat time consuming (Li *et al.*,2001).

IQR (interquartile range normalization)

Let Y_{ij} be the *j*-th probe intensity from ith slide. The IQR normalization procedure is as follows (Park *et al*, 2005):

$$Y_{ij}^{Norm} = \{Y_{ij} - med_j(Y_{ij})\}\frac{\max(IQR_j(Y_{ij}))}{IQR_j(Y_{ij})} + \max\{med_j(Y_{ij})\},$$

where $med_j(Y_{ij})$ is the median of j-th gene across all slides.

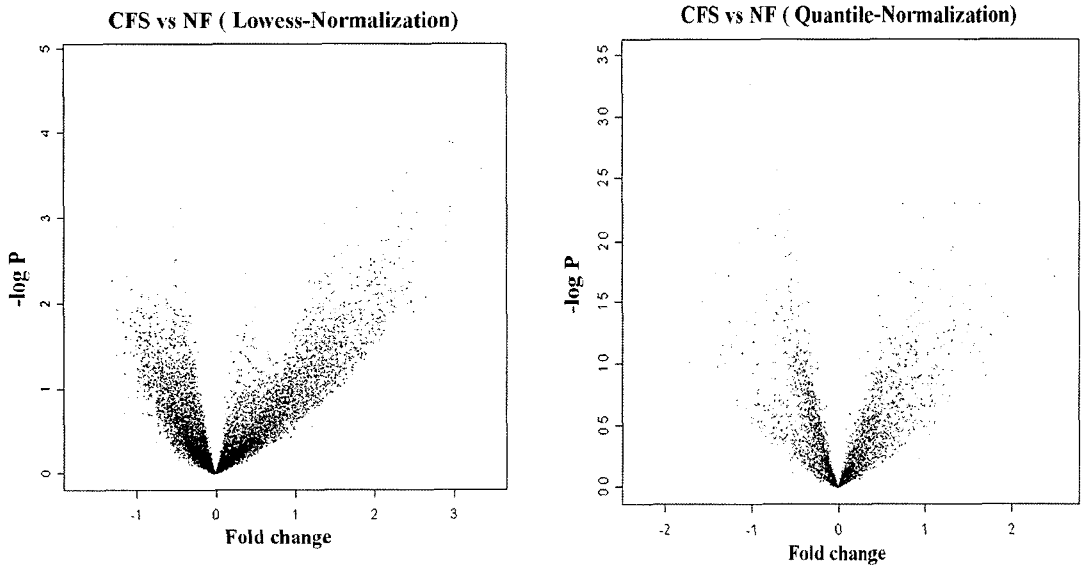
Fig. 4 summarizes the plots after applying these normalization methods. Fig. 4 shows the different effects of normalization methods. Fig 4(a) shows the patterns of the 25th, 50th, and 75th quantiles of raw data from 170 slides. Fig 4(b) and Fig 4(c) show the pattern of each of the 25th, 50th, and 75th quantiles of LOWESS- normalized data and quantile-normalized data. In the next section, we show that these normalization methods produce quite different results for identifying differentially-expressed genes.

Results and Discussion

Table 1 is the summary table for comparing different normalization methods regarding their effects on the identification of differentially-expressed genes. It shows the estimated number of significant genes for different group pairs. The simplest statistical method for detecting differential expression is the t-test for identifying differentially-regulated genes between group pairs from 5 groups, whereas the analysis of variance (ANOVA) test was

Table1. The number of significant genes dependson normalization methods

Normalization methods (Siginificance level a=0.05)		Raw data	IQR scale	Quantile	Lowess
Anova	5 groups	0	1	0	6
T-test	Control(NF) vs Case (4 groups)	0	0	1	1
	Control(NF) vs Case(CFS)	0	0	0	0
	Control(NF) vs Case (CFS-MDDm)	15	1	6	1
	Control(NF) vs Case (ISF)	0	0	0	0
	Control(NF) vs Case (ISF-MDDm)	0	2	3	0
	Control(NF) vs Case (ISF,CFS-MDDm)	2	4	0	0
Normalization methods (Significance level a=0.1)		Raw data	IQR scale	Quantile	Lowess
Anova	5 groups	0	1	1	10
T-test	Control(NF) vs Case (4 groups)	0	0	2	1
	Control(NF) vs Case (CFS)	0	0	0	0
	Control(NF) vs Case (CFS-MDDm)	122	2	20	26
	Control(NF) vs Case (ISF)	0	0	0	0
	Control(NF) vs Case (ISF-MDDm)	6	63	3	1
	Control(NF) vs Case (ISF,CFS-MDDm)	3	11	0	16



(a) LOWESS normalization (b) Quantile normalization

Fig. 5. Volcano plot for fold-change analysis after both LOWESS and quantile normalization

applied to identify differentially-regulated genes for the five groups in Table 1. We also applied some widely-used methods for testing differential expression among genes starting with the simple fold-change criteria using the volcano plot, which can evaluate both the direction and the size of an effect using intensity ratios and log-transformed p-values simultaneously. Fig. 5(a) shows that the volcano plot for LOWESS normalization comparing CFS and NF yielded asymmetric fold-change distributions, and more genes appeared to be upregulated than downregulated due to the differences of these two group-levels, whereas quantile normalization (Fig. 5(b)) produced approximately symmetric distributions. We expected approximately the same number of up- and downregulated genes. Table 1 showed that the estimated number of regulated genes in each group comparison depended on the normalization methods used. We did not see any clear difference in these data to determine which normalization method is more appropriate for finding regulated genes. We began analyzing the original data (raw data) without applying normalization methods and then compared those results with other normalization methods. The effect was most pronounced in the comparison between NF and CFS-MDDm, depending on the normalization method. We address key issues in evaluating the effectiveness of basic statistical processing steps of microarray data that can affect the final outcome of gene expression analysis that is not intended to focus on the intrinsic data underlying biological interpretation in this study. We presented some normalization methods and showed that different normalization methods yield different results. Most previous normalization methods have focused on two-channel or Affymetrix-type data;therefore they are not easily applied to our one-channel microarray data. Due to this reason, it is necessary that some modifications and slight adaptation be applied to this kind of data to have reasonable results in further studies. The use of quality measures for analyzing individual outcomes can help in estimating the reliability of final microarray study results. In particular, the study presented here showed that when the gene effect is not as large as in our example, microarray data normalization and individual processing steps have an important effect on the final outcome,especially for the identification of differentially-expressed genes. In further studies, we may consider the effects of normalization methods when the gene effect is large. It is important to test different possibilities and analyze the effects of normalization with the appropriate tools for individual processing steps. Thus, a more reliable method is required to determine an appropriate combination of individual basic processing steps for a given dataset in order to improve the validity of one-channel expression data analysis. Therefore, the overview of these effects is essential for the biological interpretation of gene expression measurements.

Acknowledgments

This work is supported by the grant from National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126).

References

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

Edwards, D. (2003). Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics* 19, 825-833.

Futschik, M. and Crompton, T. (2004). Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biol.* 5, R60.

Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* 4, 210.

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods* 31, 265-273.

Toni W., and Elizabeth R. U. Intergration of gene expression, clinical and epidemiologic data to characterize chronic Fatigue Syndrome. *Journal of Translational Medicine*

Park, T., Yi, S. G., Lee, S.Y. and Lee , J. K. (2005). Diagnostic plots for detecting outlying slides in a cDNA microarray experiment. *Bio. Techniques* 38, 463-471.

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

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.

Li, C. and Wong, W.H.(2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences* 98, 31-36.