DNA Research **12**, 1–7 (2005)

Detection and Normalization of Biases Present in Spotted cDNA Microarray Data: A Composite Method Addressing Dye, Intensity-Dependent, Spatially-Dependent, and Print-Order Biases

Shizuka Uchida,[1,†] Yuichiro Nishida,[2,†] Kenji Satou,[2,*] Shigeru Muta,[3,‡] Kosuke Tashiro,[3,‡] and Satoru Kuhara[3,‡]

School of Material Science, Japan Advanced Institute of Science and Technology (JAIST) 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292 Japan,[1] *School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST) 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292 Japan,*[2] *and Graduate School of Bioresource and Bioenvironmental Sciences, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan*[3]

(Received 29 June 2004; revised 11 January 2005)

Abstract

Microarrays are often used to identify target genes that trigger specific diseases, to elucidate the mechanisms of drug effects, and to check SNPs. However, data from microarray experiments are well known to contain biases resulting from the experimental protocols. Therefore, in order to elucidate biological knowledge from the data, systematic biases arising from their protocols must be removed prior to any data analysis. To remove these biases, many normalization methods are used by researchers. However, not all biases are eliminated from the microarray data because not all types of errors from experimental protocols are known. In this paper, we report an effective way of removing various types of biases by treating each microarray dataset independently to detect biases present in the dataset. After the biases contained in each dataset were identified, a combination of normalization methods specifically made for each dataset was applied to remove biases one at a time.

Key words: cDNA microarray; normalization; print-order bias

1. Introduction

Microarrays are widely used in laboratories throughout the world to measure the expression levels of tens of thousands of genes simultaneously in a single chip. However, due to the varieties of biases resulting from experimental protocols of the microarray, there exits a large difference between data collected by microarrays and conventional methods, which makes it difficult to evaluate the data. To remove these biases, various normalization methods have been developed.[1–8] These normalization methods remove biases that arise from variations in the microarray technology rather than from biological differences among the RNA samples or the printed probes.[6] Currently, there are three major types of biases being considered: dye, intensity-dependent, and spatially-dependent biases. The first bias to be considered is dye bias, which is caused by labeling and detection efficiencies between the fluorescent dyes used.[4] This type of bias is noticeable when scatter plots are used to draw expression levels of each gene for an organism under study. Another type of bias due to differences in dye intensities is called intensity-dependent bias, which can be detected by using an MA-plot.[8,12,13] In recent years, biases resulting from the print-tips used in the manufacturing process of a microarray are found to affect the gene expression levels of the organism under study as well.[6] This kind of bias is known as spatially-dependent bias. Recently, Smyth et al.[9] defined yet another type of bias called print-order bias, which may be caused by differences in the purity of DNA from different amplification batches or from different clone libraries.

A major problem associated with the normalization of microarray data is that not all biases are removed by one method because more than one type of bias is typically present in a given microarray dataset. In order to remove several kinds of biases, a combination of nor-

Communicated by Satoshi Tabata

* To whom correspondence should be addressed. Tel. +81-761-511-1746, Fax. +81-761-511-1149, E-mail: ken@jaist.ac.jp

† Joint First Authors.

‡ These authors performed biological experiments.

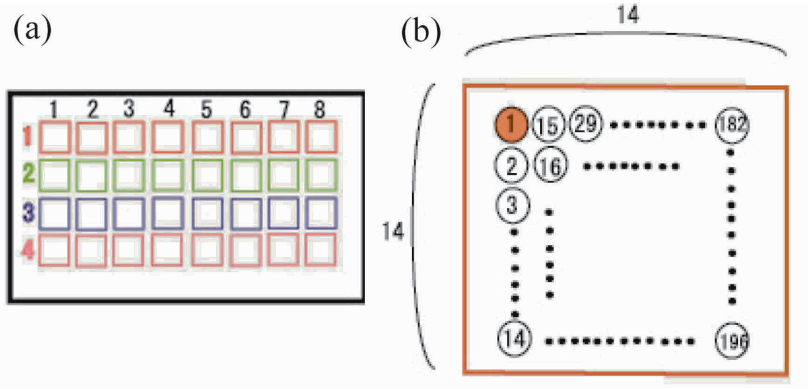

Figure 1. Alignment of the microarray chip used in this study. (a) Each row of the microarray chip is colored for visual aid. These colors correlate to the colors used in the microarray row plots (Figure 2a, b). (b) The alignment of spots for each block of the chip. The spot number corresponds to the order in which each probe was spotted.

malization methods, such as total intensity normalization for dye bias and LOWESS for intensity-dependent bias, must be applied to clean the microarray data. However, as in any kind of data transformation, inappropriate data transformations can add artifacts to the original data as by-products. In order to avoid additional biases to the microarray data, we think that it is important to understand the types of biases present in the data before applying an appropriate normalization algorithm to remove a particular kind of bias present in the majority of datasets in a particular microarray experiment.

Our method was applied to the published data for single-disruption of *Saccharomyces cerevisiae*.[10,11] First, we examined each microarray dataset using various plots and identified the specific types of biases contained in each dataset. Next, we removed biases one at a time to clean each dataset. Through this study, we propose that the application of an appropriate combination of normalization methods to a microarray dataset not only removes biases effectively but also prevents the addition of other biases to the microarray data that may obscure the true expression levels of the genes under study.

2. Materials and Methods

2.1. Microarray experiments and data

The microarray data tested were obtained from Aburatani et al.[10,11] They conducted microarray experiments to study gene regulatory networks of *S. cerevisiae*. Over 500 experimental microarray data were collected, of which 174 were shared for the analysis. For these microarray data, the same type of microarray chip was used. Figure 1 shows its alignment with eight columns and four rows. For the image analysis, two versions of the software ArrayVision (3.0 and 5.0) were used. The main difference among each version was the calculation of background intensities: either an arithmetic average for all spots on the array or the averages calculated per row (block) of the array spots. Typically, background noises are computed for each spot on the microarray, but in our case, the average or averages of background noises were calculated. Thus, it is important to consider this difference for further analysis. To counteract the lack of information concerning background noise, the raw red and green intensities were used to normalize artifacts added onto the data as a result of subtracting the same background value from all of the raw intensities in the microarray data. This, in turn, resulted in the creation of outliers for the spots whose intensity was slightly stronger than that of the background.

2.2. Methodology

Our method consists of two parts: in the first part, the different types of biases were identified from each microarray raw data; then, in the second part, an appropriate normalization method was applied to remove biases one at a time.

2.2.1. Dye bias

Dye bias is based on two assumptions: one is that the nearly equal amounts of RNAs from two samples to be compared were hybridized onto an array; another is that probes on the array were printed in a random manner to eliminate a preference towards one biological sample. When these assumptions were met and a large number of probes were spotted onto the microarray, the average intensity of Cy3 and Cy5 should be the same; in other words, the average ratio between the two dyes should equal one. Whenever this average ratio is larger or smaller than one, it is assumed that dye bias is present in an array.

To illustrate dye bias, a scatter plot was used. This plot is widely used in the microarray field. To remove this type of bias, total intensity normalization was used. Although there are many variations of normalization factors used in this type of normalization algorithm, the mean of the whole arrayed spot intensities was used to

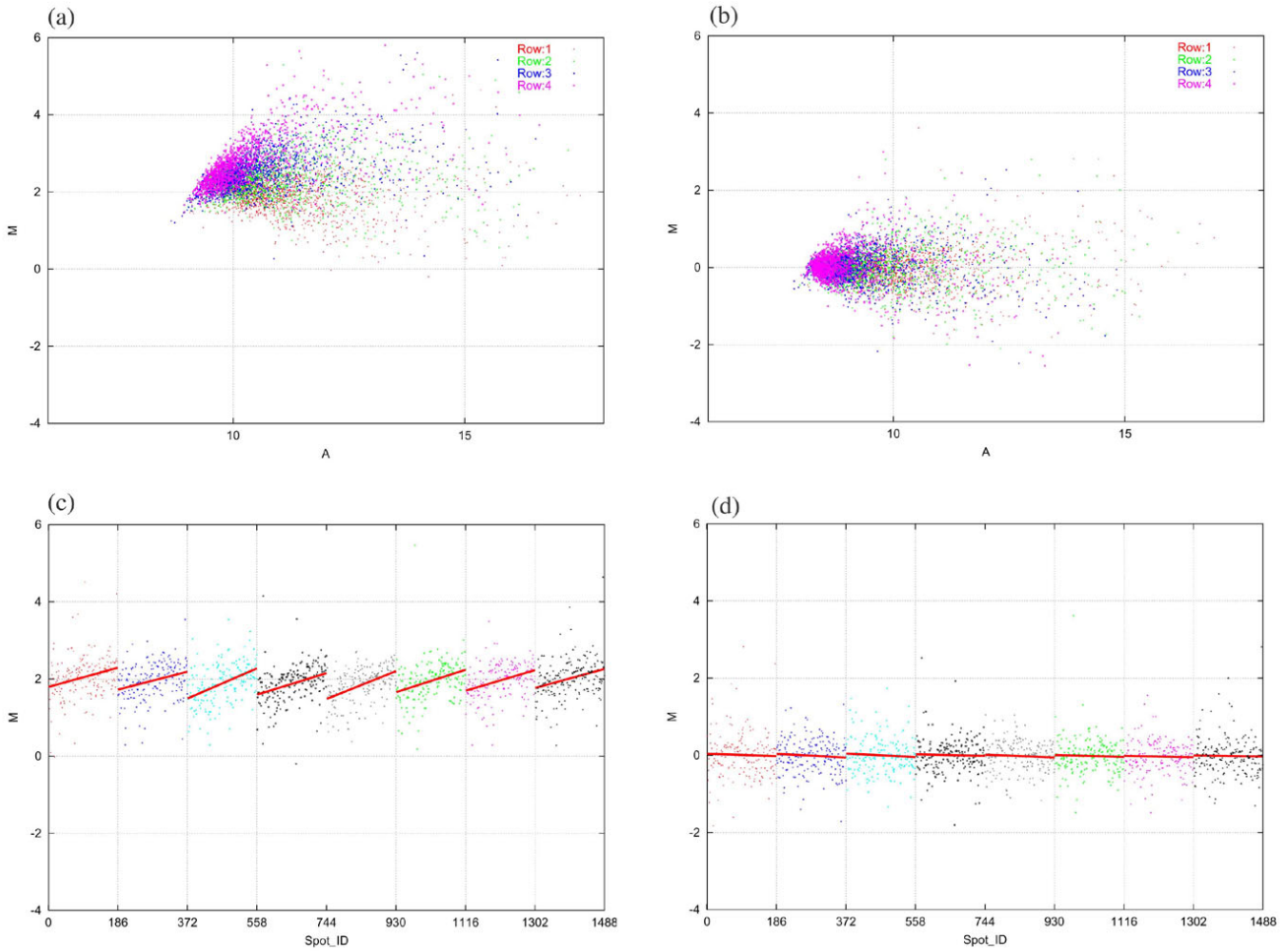

Figure 2. The microarray-row (a, b) and print-order plots (c, d) for the disruption of YMR042W. The microarray-row plot is a multicolored MA plot, in which each row of the microarray chip is colored differently to display the spatially-dependent biases according to the row of the array chip. Its x-axis is the mean log intensity for each spot, which is denoted as $A = \log_2 \sqrt{RG}$, where R is the intensity for the red dye and G is that of the green dye. Its y-axis is the log intensity ratio of dye intensities: $M = \log_2(ratio)$. Compared to (a) raw data, (b) shows how four colors assigned to each row of the microarray chip blended equally. This is the indication that the spatially-dependent bias was successfully removed. For the print-order bias, a print-order plot was used. The x-axis of this plot is the spot number according to the order in which the probes were spotted onto the array chip, and the y-axis is the M-value. A red regression line was drawn for each grid of the microarray chip. Each block contains 186 gene spots. The x-axis is the spot number according to the order in which the probes are spotted onto the array chip, and the y-axis is the M-intensity value. Notice how the slopes of the red regression lines are similar from one grid to another for (d) compared to those of (c) raw data. The plot indicates that the print-order bias was eliminated effectively.

calculate new ratios.

2.2.2. Intensity-dependent bias

An MA plot was used to reveal the intensity-dependent bias. This plot was developed by Dudoit[8,12,13] and is commonly used to show a deviation from zero for low-intensity spots.[4] LOWESS (locally weighted linear regression) was used from the LOWESS function implemented in the statistical software package R[14] to remove this type of bias.

2.2.3. Spatially-dependent bias (print-tip bias)

Compared to the above mentioned biases, this type of bias is called local because of its uneven distribution in the entire dataset.[4] It arises from inconsistencies among the spotting pins used to make the array, variability in the slide surface, and slight local differences in hybridization conditions across the array.[4] To illustrate this bias, we modified an MA plot by assigning different colors to each row of the array chip. Figure 2(a) displays typical spatially-dependent bias that was included in the microarray dataset.

Yang et al. proposed the within-print-tip-group normalization method to remove this kind of bias.[8] Their assumption was that since every grid in an array is spotted using the same print-tip, some systematic differences may exist between the print-tips used, such as slight differences in the length or in the opening of the tips, and deformation caused by many hours of printing.

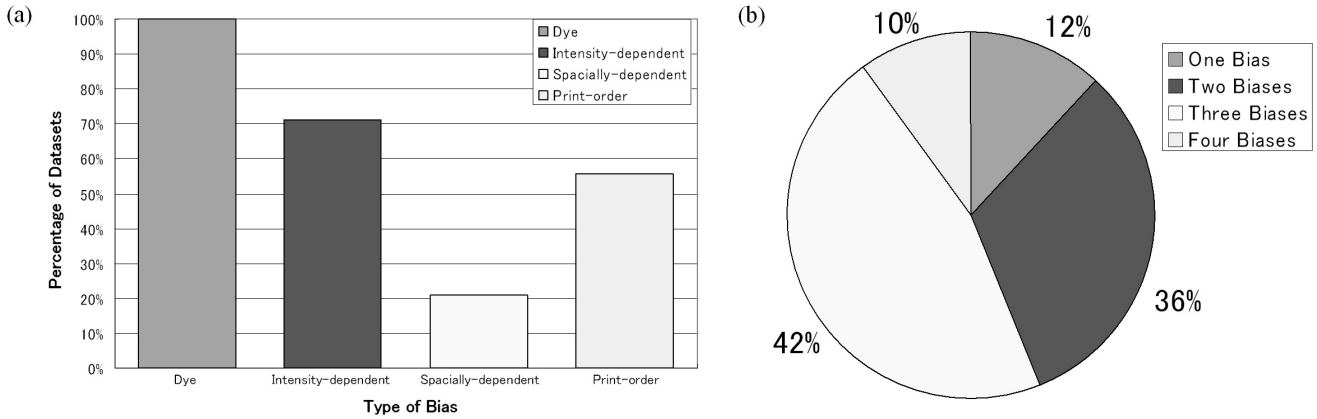

Figure 3. The graphs for (a) the number of datasets containing each bias and (b) the combinations of biases. All datasets contain dye bias, about seventy percent of the datasets include intensity-dependent bias, about a fifth contain spatially-dependent bias, and about a half includes print-order bias. Many datasets contain more than two types of biases.

The equation for this normalization is a (print-tip + A)-dependent normalization. However, unlike Yang et al., in our case, linear regression line equations were used, instead of LOWESS, to adjust M- and A-values. The following equation results:

$$M'_k = M_k - c_k(A) \tag{1}$$

where is the M'_k-value for print-tip group k after the normalization.

2.2.4. Print-order bias

There were 6272 spots present on the microarray chip used in this study. Since there were positive and negative control spots as well as empty spots on the chip, the actual number of gene spots was 5952. In order to illustrate print-order bias, these gene spots were numbered consecutively from the left top grid to the right bottom grid of the chip. Figure 2(c) shows a typical example of print-order bias. The regression lines were drawn according to the spotting pins used for the spots.

In order to detect the print-order bias, close inspections of the plots are of the utmost importance. However, interpretation of a plot may differ from one researcher to another. To minimize variations resulting from this, two conditions were set up to detect the print-order bias:

1. There are 32 regression lines drawn for each microarray chip. If the angle of the regression line is greater than 0.10 degrees, print-order bias is present.
2. If the average angle of the regression lines exceeds 0.05 degrees, print-order bias exists.

The first condition was applied to each print-tip of the array, whereas the second condition was used for the entire array. These two conditions must both be met for the detection of print-order bias.

To remove print-order bias, first, a linear regression line was drawn for the selected group of data points; then a reciprocal of its slope was multiplied by each of the data points in the group. For print-order normalization, it was assumed that the bias resulting from the order in which each probe was spotted onto the array chip correlates to the spotting pin used. With this assumption, an equation was set up as follows:

$$M'_{kn} = M_{kn} - C_k(n) \tag{2}$$

where M_{kn} is the M-value for the spot which was spotted in nth order with print-tip k and M'_{kn} is the normalized M-value.

3. Results and Discussions

Before microarray data are analyzed, normalization methods are applied to remove unwanted biases present in the data that are obscuring the true expression levels of genes. Currently, many normalization methods have been developed and applied to the microarray data. When these methods are used, raw data are transformed. However, as in any kind of data transformation, inappropriate data transformations can add artifacts to the original data as by-products. In order to avoid additional biases to the microarray data, we think that it is important to understand the types of biases present in the data before applying an appropriate normalization algorithm to remove a particular kind of bias present in the majority of the datasets under a particular microarray experiment. The biases often contained in the microarray data are dye, intensity-dependent, and spatially-dependent biases. Among these biases, we put a special emphasis on the print-order bias. Previously, this bias resulting from the order in which the spots were laid down during the printing of the array was thought to affect the data to a lesser degree compared to the intensity- and spatially-dependent biases.[9] However, Fig. 3(a) indicates that this type of bias was detected in a large proportion of datasets. Thus, we hypothesized that the removal of

this type of bias, as well as other biases, will improve the quality of microarray data.

According to Fig. 3(b), more than half of the datasets in this study included three or more biases. This indicates the importance of identifying the types of biases included in the data before applying normalization methods. To achieve this goal, we used various graphical representations to detect the types of biases present in each microarray dataset and then applied appropriate normalization methods to remove biases one at a time. As explained in the Methodology section, total intensity normalization was used to remove the dye bias, LOWESS for the intensity-dependent bias, and linear regression equations for the spatially-dependent and print-order biases. The order in which various normalization methods were applied is crucial. If a dataset contains all of the biases, it is the best, from our experience, to remove these biases in the following order: dye, print-order, spatially-dependent, and intensity-dependent. This sequence is due to the fact that all biases except the intensity-dependent are removed in a linear manner, whereas the intensity-dependent bias is removed by LOWESS, which transforms the dataset in a non-linear manner and greatly depends on the smoothing parameter (0.33 was used in our case). Therefore, if LOWESS is applied prior to other normalization methods, other kinds of artifacts might be added to the dataset due to its inappropriate data transformation. Due to the purpose of data mining, treating microarray datasets independently is not a common method. However, we argue that each microarray chip does contain unique biases that differ from one dataset to another; therefore, normalization methods must be customized to remove biases present in each dataset.

Figure 2 (b) and (d) illustrate how the typical dataset that included all four types of biases identified in this study were normalized through our stepwise normalization methods. As noticeable from the figure, all biases were effectively removed.

Currently, there are many normalization methods available to remove unwanted biases. The main assumption of these normalization methods is that most genes are non-differentially expressed. This is particularly true for the data our method was tested on because only one gene was disrupted in each target sample. Therefore, we speculate that only a small percentage of total gene spots show significant expression changes, and thus the standard deviations of these datasets must be small if and only if artifacts obscuring the true gene expression levels are removed appropriately. Table 1 shows how our method successfully reduced the average standard deviation for the 173 datasets (excluding one control experiment) used in this study. For the application of all normalization methods, normalization methods were used in the following order: dye, print-order, spatially-dependent, and intensity-dependent. However, compared to the application of all normalization methods, the average standard deviation for our method was larger. Therefore, in order to show the effectiveness of our method, our method and the application of all normalization methods were compared based on two criteria: skewness and kurtosis. Skewness is a measure of symmetry for the data distribution. It is calculated by $\sum_{i=1}^{N}(Y_i - \overline{Y})^3/s^3$, where $\overline{Y}$ is the mean, s is the standard deviation, and N is the number of data points. The skewness of a normal distribution, which looks the same to the left and right of the center point, is zero. Kurtosis is a degree of the relative peakedness of a distribution compared to the normal distribution. The kurtosis for a standard normal distribution is three. For this reason, in many cases, the following formula is used for calculating kurtosis: $(\sum_{i=1}^{N}(Y_i - \overline{Y})^4/s^4) - 3$, where $\overline{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Thus, this formula is used here as well. Table 2 shows how our method outperformed the application of all normalization methods based on the abovementioned statistical measures. Figure 4 displays the distributions

Table 1. The average standard deviations after the application of normalization methods for the 173 datasets used in this study. Since dye bias was present in each dataset, total intensity normalization was applied prior to applying other normalization methods.

Normalization Method	Average Standard Deviation
Total Intensity Normalization	0.3669
LOWESS	0.3138
Microarray-Row	0.3080
Print-Order	0.3177
All Normalization Methods	0.2621
Our Method	0.2753

Table 2. Comparison between our method and application of all normalization methods. If the skewness of the dataset after the application of our method was better, it was counted as yes; whereas if the one after the application of all normalization methods was better, it was counted as no. The same rule was used for kurtosis. Both Yes indicates that the application of our method was better in both criteria. In the case of Yes-No, as shown in the table, means that the skewness of the dataset after the application of our method was better; however, the kurtosis was not better than the one after the application of all the normalization methods. Equal means that all normalization methods were applied for our method.

	vs All Normalization Methods
Both Yes	74
Both No	19
Equal	16
Yes-No	10
No-Yes	54
Total	173

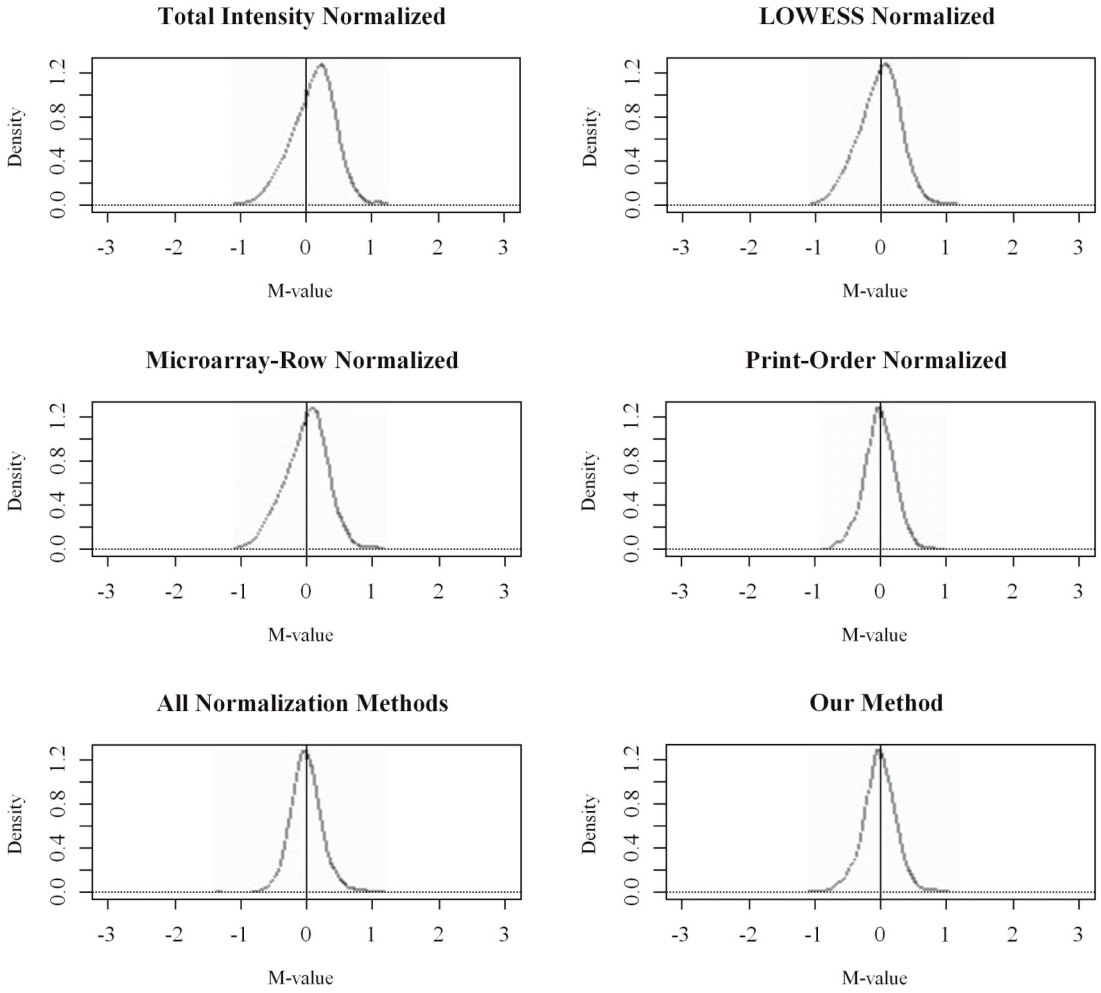

Figure 4. Graphs showing the density distributions of the dataset for the disruption of YEL056W after the applications of various normalization methods. Since dye bias was present in each dataset, total intensity normalization was applied prior to applying other normalization methods. The x-axis is the M-intensity value and the y-axis is the density distribution of the dataset. Each normalization method is indicated in the title of the graph. The dataset shown here contains dye and print-order biases.

after the application of various normalization methods. The dataset shown in the figure contains dye and print-order biases. At glance, the distribution after the application of our method does not seem to differ from the one after the application of all normalization methods. However, the skewness and kurtosis for the normalized data by our method were 0.0970 and 3.3409, respectively, whereas those for the one after the application of all normalization methods were 0.5790 and 4.3317, respectively. The difference seems minor, but it might result in the misidentification of differentially expressed genes. Thus, caution must be taken before applying currently available normalization methods *a priori*.

The cleanliness of the original data is the most important factor in elucidating the biological information contained in the microarray data. However, not all of the microarray data can meet this condition. Therefore, normalization methods are necessary. Although normalization methods only help to clear the gray parts of the data that are obscuring the true expression levels of genes, we propose that the application of an appropriate combination of normalization methods to a microarray dataset not only removes biases effectively but also avoids additional biases to be added onto the microarray data.

Acknowledgements: The authors would like to thank Dr. Robert DiGiovanni and Ms. Petra Ruick for grammatical corrections. This work was supported by JSPS research for the future program, a grant-in-aid for Science Research on Priority Areas from the MEXT, and a grant-in-aid from Institute for Bioinformatics Research and Development, Japan Science and Technology Corporation (BIRD-JST).

References

1. Baldi, P. and Hatfield, G. W. 2002, DNA microarrays and gene expression, Cambridge University Press, Cambridge, United Kingdom.

2. Hatfield, G. W., Hung, S., and Baldi, P. 2003, Differential analysis of DNA microarray gene expression data, *Molecular Microbiology*, **47**(4), 871–877.
3. Kerr, M. K., Martin, M., and Churchill, G. A. 2000, Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, **7**(6), 819–837.
4. Quackenbush, J. 2002, Microarray data normalization and transformation, *Nature Genetics Supplement*, **32**, 496–501.
5. Rocke, D. M. and Durbin, B. 2001, A model for measurement error for gene expression arrays, *Journal of Computational Biology*, **8**(6), 557–569.
6. Smyth, G. K., Yang, Y. H., and Speed, T. 2002, Statistical issues in cDNA microarray data analysis. In Brownstein, M. and Khodursky, A. (eds.), *Functional genomics: methods and protocols*, Methods in Molecular Biology Humana Press Totowa, NJ.
7. Wolfinger, R. D., Gibson, G., Wolfinger, E. D. et al. 2001, Assessing gene significance from cDNA microarray expression data via mixed models, *Journal of Computational Biology*, **8**(6), 625–637.
8. Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. 2001, Normalization for cDNA microarray data. In Bittner, M., Chen, Y., Dorsel, A., and Dougherty,E. (eds.), *Optical Technologies and Informatics*, San Jose, CA: International Society for Optical Engineering.
9. Smyth, G. K. and Speed, T. 2003, Normalization of cDNA microarray data, *Methods*, **31**(4), 265–273.
10. Aburatani, S., Tashiro, K., Savoie, C. J. et al. 2003, Discovery of novel transcription control relationships with gene regulatory networks generated from multiple-disruption full genome expression libraries, *DNA Research*, **10**, 1–8.
11. Kuhara, S., Tashiro, K., and Muta, S. 2002, Drug discovery based on microarray, *Nippon Yakurigaku Zasshi*, **120** (Suppl. 1), 47–50 [Article in Japanese].
12. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. 2002, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, **12**, 111–139.
13. Yang, Y. H., Dudoit, S., Luu, P. et al. 2002, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, **30**(4), e15.
14. Ihaka, R. and Gentleman, R. 1996, R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.