cDNA microarray data normalization

Mohammad Mehdi Naghizadeh, Ebrahim Hajizadeh* and Anoshirvan Kazemnejad

Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modarres University, P.O.Box: 14115-331, Tehran, I.R. Iran

Abstract
Normalization process is a set of operations whereby systematic biases are removed from microarray data. Therefore researcher can attain acceptable results and have more logic comparisons. This process considering the bias constructions and the effects on microarray is recognizable and applicable. Examples of this process are spatial correction, background correction, rescaling, dye effect and within slide normalization. Spot intensity standard deviation is decreased by that normalization processes, and confidence to the microarray data analysis outcomes is increased. This paper describes all of these methods. In addition it contains the examples of a real dual channel cDNA microarray experiment to illustrate normalization process.
Keywords: *Microarray; Normalization; cDNA; Preprocessing; Background correction; Robust smoother.*

INTRODUCTION

Microarray technology, as a measurement method, has enabled researchers to study thousands expression of genes simultaneously (DeRisi *et al*., 1996). Like other measurement methods, microarray technology is not free of systematic biases. These may occur while labeling, spotting, scanning, image analysis or other similar operations. But these factors in no way add to biological differences (Schena, 2000). There are different methods to create a microarray. The first step is to extract RNA from intended tissues. These tissues are usually samples of groups under study, control group and experiment group. Second step is reverse transcription and formation of cDNA. In this way two types of cDNA are formed. These two types of cDNA are labeled with different dyes. In new technologies, two types of fluorescent namely, Cy3 and Cy5 are used. These two types of fluorescent are detectable by green and red laser (Pat Brown Laboratory Protocol, 2002).

Then labeled cDNAs are mixed and spotted as targets on a slide whereon are put certain probes of nucleotide sequence. In the next step that is called hybridization, cDNAs are provided with such opportunity to hybridize with probes. The slide is washed out at the end of this stage so that those groups of cDNAs that have not hybridized are removed. At the end, the slide is scanned with two 532 nm and 635 nm lasers, and then two green and red images are produced. Different Image Analysis methods create a matrix by merging these two images, where each element of this matrix shows the intensity of green or red dyes at a spot (Stekel, 2003).

Microarrays are infected with types of systematic biases considering their natures, such that if a microarray experiment is conducted twice with the same previous materials and conditions, then genes expression may have different quantities after image scan and analysis (Yang and Speed, 2003). In brief, the following resource can be pointed out in creating bias microarray experiments: including biases resulting from preparation, RNA extraction and transcription, difference in labeling, bias resulting from chemical structure and slide heterogeneity, problems related to spotting, factors creating bias in hybridization, errors resulting from scanning and different image analysis methods and data quantification (Draghici, 2003). The set of these factors create disasters while comparing the genes to such extent one can not find if expression

Correspondence to:* **Ebrahim Hajizadeh, Ph.D.
Tel: +98 21 8011001; Fax: +98 21 8013030
E-mail: Hajizadeh@modares.ac.ir

of this gene is due to genetic properties or due to systematic biases. Hence, it seems necessary to set forth subjects related to normalization to eliminate or lessen effects of these biases.

Normalization process is a set of operations whereby systematic biases are removed from microarray data so that researchers could reach more proper results and have more logical comparisons in the study of different genes expression (Xiao *et al*., 2004). Like other bioinformatics processes, microarray technology consists of two laboratory sections and computing section; and, due to complexity of its computing section it is somehow ambiguous for biologists. Moreover, since production technology of this new-emerging phenomenon has recently entered into the country, and since with suitable and proper design one can reach acceptable and reliable answers in these experiments hence, conducting this research with the aim to explain and elaborate normalization process in microarray technology seems to be necessary. In this article, after a brief review of how image is analyzed, we will discuss bias removing methods in microarray data, and will show above stages. We hope that this article will pave the way for the extraction of proper results based on a proper design of a microarray test.

MATERIALS AND METHODS

Image Analysis: After laboratory stages, hybridization, and washing, microarray slides are ready for analysis. Each slide is presumptively divided into thousands of pixels considering power and precision of scanner. Laser is radiated on each pixel on two channels, where reflection of which produces different heat on sensors.

This heat as the intensity of red or green dyes is quantified and is changed into numerical values. Each spot consists of hundreds of pixels. This means that thousands of values are reported for each spot. Spot dye intensity, is a function of above values. In the first choice, the *mean* of these values can be considered as spot dye intensity. But the *median* is a better choice due to non-sensitivity to outlier data. For this reason, this landmark is used in most microarray tests. Nunez-Garcia *et al*. has suggested *mode* as the more accurate statistic in this regard (Nunez-Garcia, 2004).

Scanner sweeps both the spots and their surroundings (so called as background), and reports the amounts of red or green dye intensity existing in that area. Dye intensity in this area is called as "background dye intensity". Scanner reports too many numbers for the background which could be summarized as the mean, median, or other functions. It is better that the function of the background intensity be similar to the function of foreground intensity (Kamberova and Shah, 2002). Figure 1 shows general view of a spot and different methods for calculation of background dye.

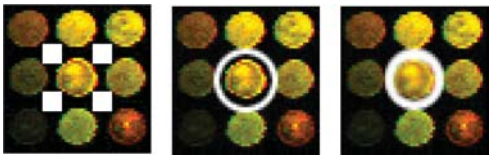

Figure 1. Inside bright domain are called foreground (pixel) and inside shaded part are called background. a, b and c show three methods for calculation of back ground dye.

Spatial Correction: Spotting process on slides may last for several hours during which final spots may be less coordinated with other spots due to dryness of spotter's pins. Such disorder may occur as a result of hybridization or some other technical problems in image analysis. To correct such bias that exists on the border of each slide in particular, and supposing spatial (location) effect is a multiple of dye intensity and is equal on both channels, and let $S_R(x,y)$ as dye intensity of red channel at (x,y), $F(x,y)$ is real red dye intensity at (x,y), and $C(x,y)$ is the spatial (location) effect that is equal on both channels and is a multiple of real dye intensity, the following method is suggested. If proportion of red and greed dye intensities are used in the equations instead of red and greed dye intensities, this proportion shall be free of any spatial (location) effect:

$$\frac{S_R(x,y)}{S_G(x,y)} = \frac{F_R(x,y)C(x,y)}{F_G(x,y)C(x,y)} = \frac{F_R(x,y)}{F_G(x,y)}$$

This means that signal dye intensity ratio is equal to true dye intensity ratio (Wit and McClure, 2004). If the above-mentioned presuppositions are not approved, we can use two-way smoothing which is widely discussed in Stekel's book (Stekel, 2003).

Background Correction: Since no spot was spotted in the background area and therefore red and green dye intensities should be zero, actually it was not so. Therefore, in any can conclude that both the red and green channels have reported a value higher than the true intensity. Common method to correct background in microarray experiments based on this presupposition that background dye intensity is additive to main dye intensity and based on this assumption, to calculate dye intensity it is enough to subtract background

dye intensity value from main dye intensity value (Kim *et al*., 2002).

Corrected dye intensity = reported dye intensity (signal) – background dye intensity

$$R_i = Rf_i - Rb_i$$
$$G_i = Gf_i - Gb_i$$

To correct background intensity, other suggestions have been proposed. As it was told before, each spot consists of hundreds of pixels and finally a function of pixels dye intensity is considered as dye intensity of each spot (usually median). By subtracting three standard deviations of background dye intensity from dye intensity function, the corrected value is calculated (Suite, 1999):

Corrected dye intensity = reported dye intensity (signal) - 3 standard deviations of background dye intensity

$$R_i = Rf_i - 3Se(Rb_i)$$
$$G_i = Gf_i - 3Se(Gb_i)$$

As it was discussed before, both above-mentioned patterns are based on this assumption that background dye intensity is an additive quantity. This presupposition simplify the calculations but creates other problems, for example consider a case where the reported dye intensity (signal) is less than background dye intensity, and main dye intensity that is gained by subtracting these two values is unavoidably less than zero.

To avoid such problem we can remove such spots as flagged biases from analysis (Kim *et al*., 2002). But *Bakewell* and *Wit* believe that removing such spots will result in loss of information (Bakewell and Wit, 2004). For this reason, they weighted each spot with inverse dye intensity deviation, and minimized the effect of biased spots in this way. Another model called probabilistic model was presented by Irizarry *et al*. (2003). In this method, conditional expectation of main dye intensity is considered as the basis of calculations given that signal dye intensity values and background dye are known.

E(corrected dye intensity | signal dye intensity + background dye intensity)

$$E(R_i \mid Rf_i + Rb_i)$$
$$E(G_i \mid Gf_i + Gb_i)$$

Where distributions of $Rf_i + Rb_i$ and $Gf_i + Gb_i$ are supposed to be normal. Calculation formulas of this model are not presented here due to complexity of probability subject. For more study Irizarry is suggested for reference (Irizarry *et al*., 2003). The most advantage of this model is that there are less flagged spots in this model.

Rescaling: Values calculated for green and red dyes after background correction are usually skew or abnormal distribution which will result in doubtful statistical interpretations. To avoid such problem, the best suggestion is to change measurement scale and using a logarithm scale. Using logarithm scale gives such feature to the data that values changes from multiplication mode to addition mode. For example suppose following numbers *x*, *2x* and *4x*, if binary logarithm of these numbers are applied instead of the actual values, values will change as follows:

$$\log_2 x,\ \log_2 2x = \log_2 x + 1,\ \log_2 4x = \log_2 x + 2$$

for each time that *x* is doubled, one unit is added to logarithm scale. For this reason scale of logarithm measurement is improved in binary format (Quackenbush, 2002). After rescaling, following equation is used for the data. For spots $i = 1,2,\ldots, p$, green and red dye intensity shall be R_i and G_i, and for these values:

$$M_i = \log_2(R_i/G_i) = \log_2 R_i - \log_2 G_i$$

$$A_i = \log_2\sqrt{R_i G_i} = \frac{1}{2}(\log_2 R_i + \log_2 G_i)$$

By drawing these two values beside each other, the result shall be *MA*-plot that is a 45 degree rotation of $\log_2 R$ against $\log_2 G$.

Dye Effect Normalization: Main presupposition in the microarray experiment is that both types of cDNA are hybridized and joint with probes similarly. Nowadays, microarray experiments based on dye intensity measurement of those cDNAs that is are labeled with two Cy3 and Cy5 fluorescent. Contrary to similarities between these two dyes, they are in some way different from each other including size of molecules. Due to larger size of Cy5 molecule, in the middle of the diagram, data is located in a lower limits compared to two ends that associates *banana effects*. This problem may occur due to image analysis defects. Banana effect may result in a darker dye in one channel that the other. It means that one cDNAs expressed more than other. Above presupposition is true in all components of a microarray slide. This means that if the slide is divided as per pins of spotters, green and red dye intensity should be equal in all subgroups.

To avoid such error, Yang *et al*. (2001) set forth dye effect normalization. Such that M and A values that were previously discussed here in this article, are cor-

rected as follows:

$$\hat{A} = A$$

$$\hat{M} = M - f_i(A)$$

Supposing that intensity of both dyes are proportionate with each other $R=kG$, then by selecting mean or median of logarithm ratio to:

$$f(A) = mean(M) = mean(\log_2(R/G))$$

And with subtracting it from M, global normalization has been done.

Moreover, by fitting *loess* or *lowess* regression function between A and M, and calculating remainders resulting from subtraction of estimated values of regression function from M, and even separately and on classification of pins, a more accurate normalization is carried out.

Within Slide Normalization: After dye effect normalization, logarithm ratio of both dyes shall be concentrated around zero as per classification of pins of spotter. But this does not mean that they distribution of these pins is equal. Exactly for this reason, within slide normalization shall be needed. Supposing that the entire logarithm of dyes proportion follows normal distribution with zero deviation and $\alpha_i^2\sigma^2$. It shall be enough to estimate α_i^2 to equalize distribution of all logarithms. For this reason, *Yang et al*. (2002) suggest the following estimate:

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}}$$

$$MAD_i = median_j\left\{\left|M_{ij} - madian(M_{ij})\right|\right\}$$

Outstanding feature of MAD (Median Absolute Deviation) is that this statistic is not affected from outliers. In this way, gene expressions are not so affected.

Data: Data used here in this paper is derived from the Bullinger *et al.* (2004). That study aimed to review gene expression in adult patients with acute myeloid leukemia. For this purpose, 65 samples of cDNA of peripheral blood and 54 samples of bone marrow of 116 adult patients with AML (Acute Myeloid Leukemia) were extracted of which karyotype of 54 persons was normal. Microarray slide formed after hybridization by *Genepix 4000B* scanner was quantified. This data is available at Stanford university microarray database (Gollub *et al*., 2003).

For the purpose of data normalization in this study, *R* software and *maARRAY* and *Smida* package were used. *maARRAY* is explained in *Dudoit* (Dudoit and Yang, 2003), and Wit's book has more information about *Smida* (Wit and McClure, 2004). All of two packages are available at Bioconductor web site: Http://www.bioconductor.org. *Smida* also can be downloaded form:
http://www.stats.gla.ac.uk/~microarray/book/**.**

RESULTS

Average and standard deviation of pixel intensity means are 476.2 ± 671.1 in green channel and 372.6 ± 599.0 in red channel. 1508 spot from 43200 spot (3.5 percent) are greatest intensity than average plus three standard deviation ($m + 3s$) in both channels. Average and standard deviation of pixel intensity median are 454.7 ± 626.6 in channel one and 355.9 ± 545.0 in another channel.

Among the set of data existing after quantifying image of microarray slide, spot pixels intensity median was selected as signal dye intensity. Because this quantity has more centralize around its average, and for such reason background pixels dye intensity median was selected as background dye intensity to be applied.

Figure 2 is the spatial diagram of green channel background intensity. Since there is no target in the background to joint with probes, therefore, dye intensity should be uniform throughout the diagram. But if it is observed in figure 2 that dye intensity at the edge and around column 3 and row 6 is more than other areas. Non-uniformity of background dye is a sign of presence of spatial effect (location) in the microarray. To remove the spatial effect, spatial bias correction should be applied. If background intensity spatial diagram of other channel is drawn, we will have a model similar to the first diagram. Equal model of two diagrams shows that spatial effect is equal in two channels. Moreover, it is not unlikely that spatial effect is a multiple of true dye intensity. By applying these two hypotheses we may use the calculated dye intensity to eliminate this effect. For this reason, we applied proportion of these two channels during normalization process, and as it can be observed in spatial diagram after preprocessing (Fig. 3), dye distribution seems to be smoother in morphological point view. This diagram is drawn after normalization process and based on $M=\log_2(R/G)$.

As it was mentioned before, elimination of background effect is the most important phases of normal-

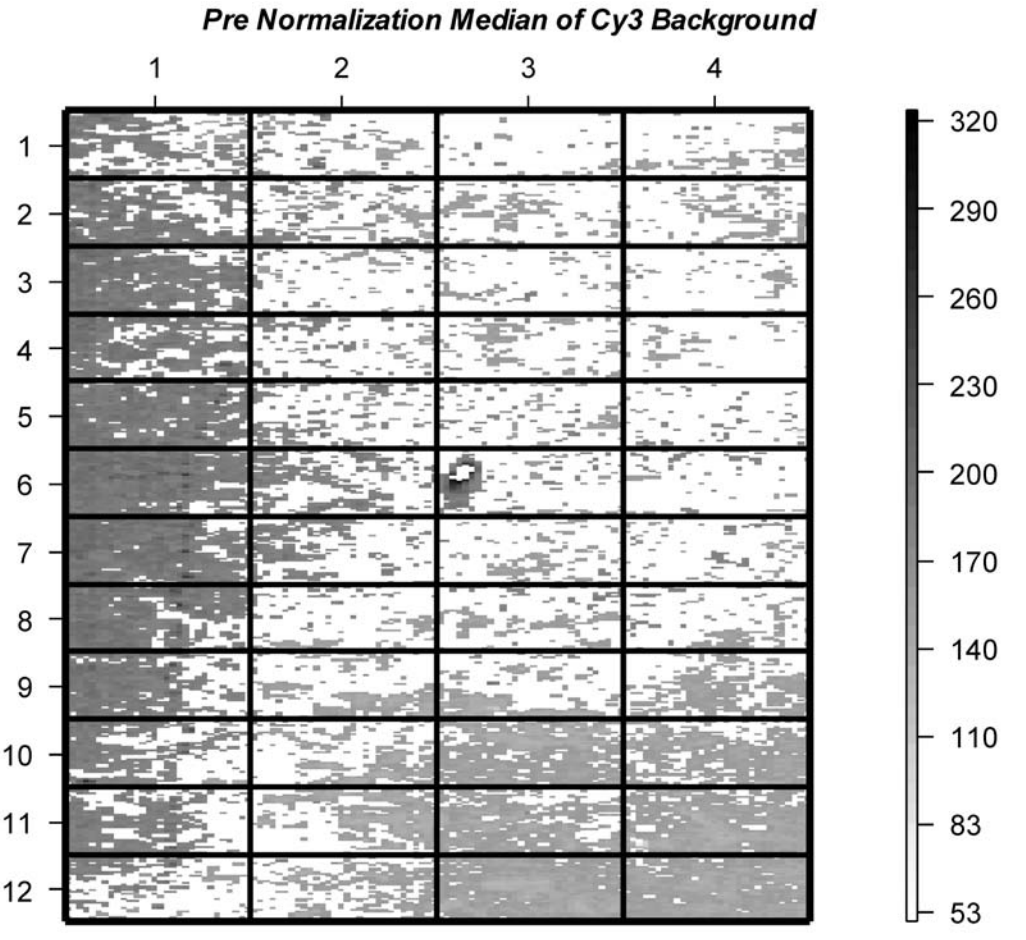

Figure 2. Spatial plot of the morphological green channel background intensity. Dye intensity at the edge and around column 3 and row 6 is more than other area.

ization process. To neutralize this effect as it is common, we supposed background dye effect as a value additive to signal dye intensity value. By subtracting these two values from each other, we tried to separate background effect from signal. Figure 4-left is a scoter plot that shows green and red dye intensity values after subtracting background dye. The most outstanding factor that can be seen in this plot is the presence of spots that are less than zero in one of the two axes. The reason is that signal dye intensity is less than background dye intensity in these spots. For this reason, such spots are called biased spots, since we have to put them aside from analysis, but if we use background correction with probabilistic model, number of biased spots shall be considerably less. If the case is clearly shown in the plot, then in figure 4-left, we have used background dye correction with probabilistic model.

Histogram of spots' green channel dye intensity is shown in figure 5-left. If we can observe in this diagram that data have skew distribution, we applied binary logarithm to eliminate such problem. As it can be observed in figure 5-right, data distribution is more symmetric after logarithm process and it is less skew.

Standard deviation of logarithm of both two channel intensity are 2.56 that has become less than standard deviation of intensity before this stage. Number of spot with intensity lowest than mean plus three standard deviation ($m + 3s$), has become zero, and just 392 spot have intensity more than mean plus two standard deviation ($m + 2s$) of all spot intensity. It means that the data has become stable after taking logarithm.

In previous section, we pointed that in a microarray experiment, both types of cDNA are equally joint with probes. We expect that green and red dye intensity is uniform throughout the slide. If such presupposition is true, data is distributed equally around MA-plot. While data distribution is not equal in figure 6-left, to eliminate such problem, loess regression function of first class degree with *span=0.3* we fit each pin with data. The result is shown in figure 6-right.

Data is distributed around zero in diagram 6-right.

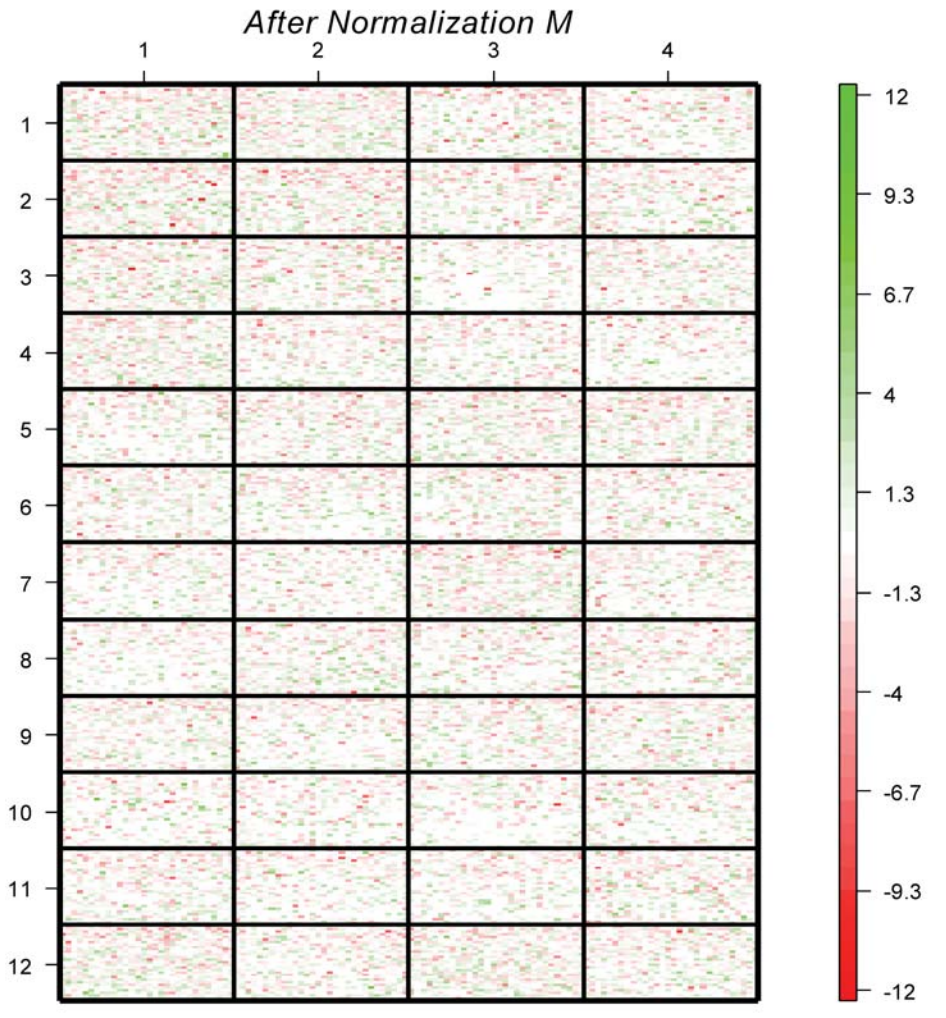

Figure 3. Spatial plot of two channel intensity log-ratio using green to red color palette. Green means that sample genes are more expressed and red means that reference genes are expressed more. This plot was drown after normalization and two dye distributions are uniform in all around.

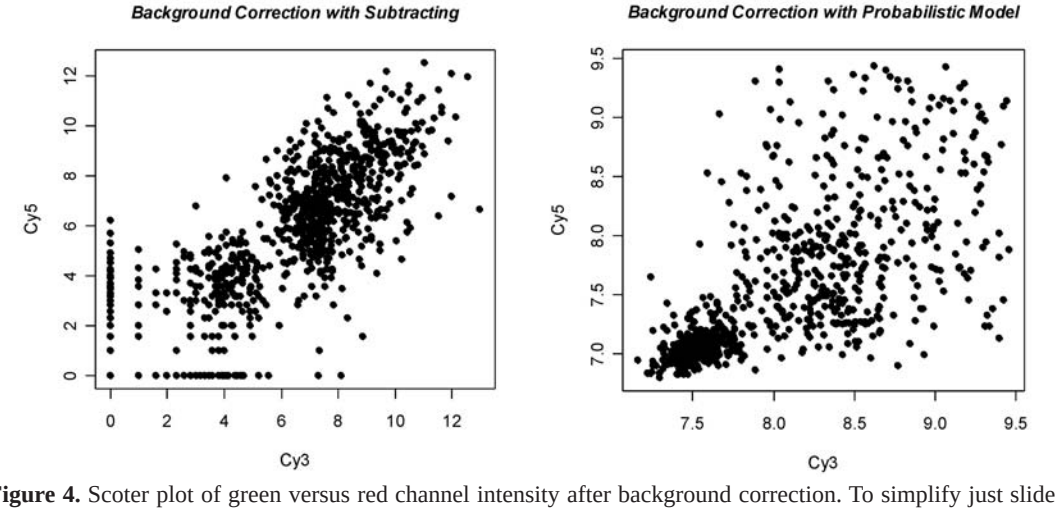

Figure 4. Scoter plot of green versus red channel intensity after background correction. To simplify just slide one is shown. The process is made simplify by presupposition that background intensity is additive to pixel intensity. But it make flagged data problem. In left plot data near the axes is flagged. In right plot this problem was solved by using of probabilistic model.

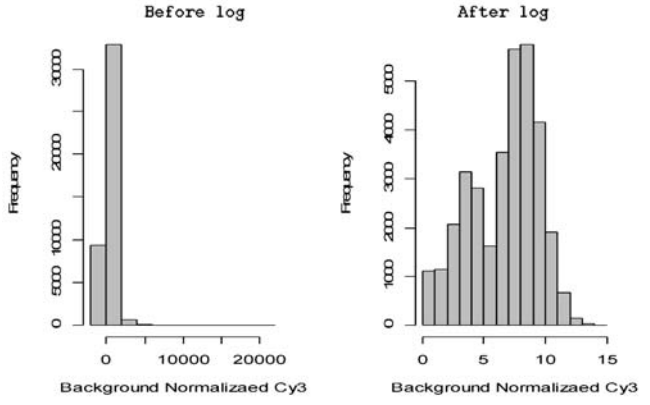

Figure 5. Green channel histogram after background correction, before taking logarithm (left) and after it (right). Data distribution is changed to normal by logarithmic transformation.

But does this distribution is equal as per spots' locations? The answer is shown in figure 7. Box plot which is drawn before dye effect normalization, data average is not zero. In figure 7-left the box plot is drawn after dye effect normalization wherein data mean is clearly equal to zero. But data distribution is not equal as per spots' locations classification. That is boxes' lengths are not equal. To eliminate such problem and prepare data we used within slide normalization to extract statistical interpretations that needed variance equality. Box plot of figure 7-right was drawn after such normalization that shows data distribution equality. After end stage normalization average and standard deviation of green and red channel are 6.74 ± 2.50 and 6.78 ± 2.52 respectively. Percent of spot that have intensity more than mean plus two standard deviation of all spot intensity has become less than before of this stage (Table 1). This shown that the data has become more stable after normalization processing. Finally about 31126 spot are useable in next microarray data analysis such as clustering.

Now, data of this microarray experiment are ready for statistic analysis and interpretations. Final corrected dye intensities of R and G are calculated by doing a few mathematical transformations.

DISCUSSION

Normalization is one of the main phases in a microarray experiments, because precision of the results is increased by providing interpretation ground. In this article, we have tried to introduce existing normalization methods and discuss about the necessity of using these methods. Data of the study conducted by Bullinger *et al*. (2004) was practically normalized.

Based on the presented subjects, we can conclude that considering several sources for systematic biases in microarray experiment, normalization process is unavoidable. For example, in above data, considering non-uniformity of data distribution throughout the slide, spatial effect is considered to be necessary and due to similarity of this effect in these two channels, spatial effect was neutralized by calculating ratio of dye intensity of both channels.

If additive models are used (Kim *et al*., 2002), we might have faced a great quantity of biased data. For this reason, we used probabilistic model to eliminate the background effect. Due to high dye intensity vari-

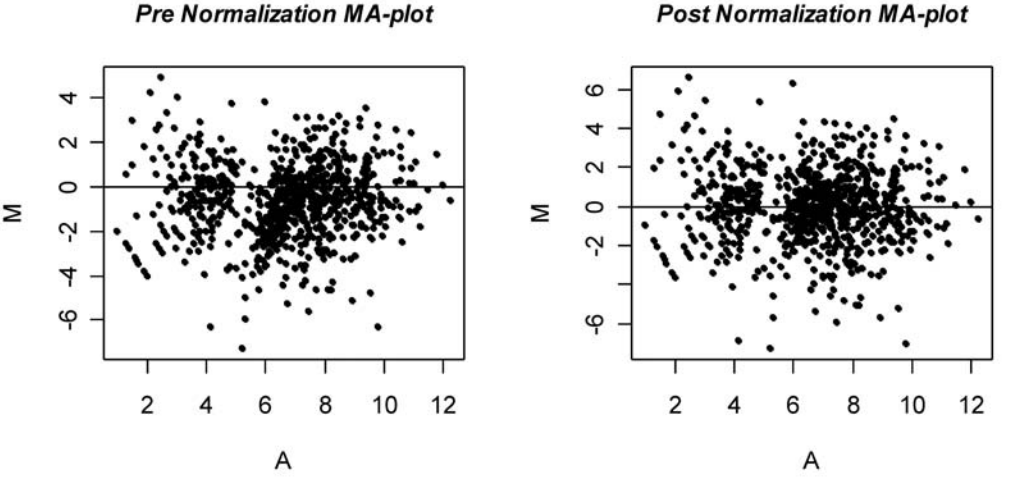

Figure 6. MA plot of slide one of data before (left) and after (right) fitting loess regression. Data is cauterized around zero by fitting loess regression.

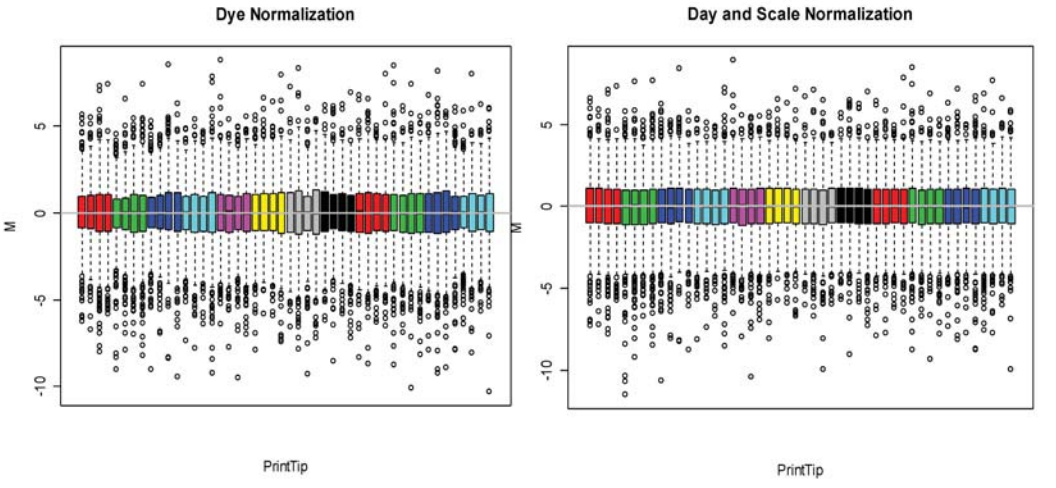

Figure 7. Box plot of two channel log-ratio per location. Before with in slide normalization (left) mean of data is zero. After it (right) all of variance is also equal. The boxes length is equal.

Table 1. Mean, Standard deviation (Std), Number and percent of spot with greatest intensity than mean plus two and three standard deviation in green and red channel at different stage of normalization.

Stage	Chanal	Mean	Std	>M+2s	%>M+2s	>M+3s	%>M+3
Before normalization	Green Net	365.0	683.6	4178	12.40	603	1.79
	Red net	378.6	583.4	4802	14.26	712	1.99
Background normalization and taking logarithm	Log Green	6.7	2.6	194	0.62	0	0
	Log Red	6.8	2.6	198	0.64	0	0
Local loess and scale normalization	Log Green	6.7	2.5	181	0.58	0	0
	Log Red	6.8	2.5	191	0.61	0	0

ance in this study compared to other microarray tests, banana effect was not clear though the mean was not zero. For this reason, first degree loess regression is used to eliminate dye effect (Yang *et al*., 2001). Variance difference has been eliminated with other method proposed by Yang (Yang *et al*., 2002). Spot intensity standard deviation is decreased by preprocessing processes (Table 1) and confidence to the microarray data analysis outcome is increased.

References

Bakewell DJ, Wit E (2004). Weighted analysis of microarray gene expression using maximum likelihood, *Bioinformatics*, 2: 229-239.

Bullinger L, Döhner K, Bair E, Fröhling S, Schlenk RF, Tibshirani R, Döhner H, Pollack JR (2004). Use of Gene Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. *New Engl J Med*. 350:1605-16.

DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet*. 14: 457–460.

Draghici S (2003). *Data Analysis Tools for DNA Microarray*, Chapman & Hall/CRC Press.

Dudoit S, Yang YH (2003). Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In: *The Analysis of Gene Expression Data: Methods and Software*. Parmigiani G, Garrett ES, Irizarry RA, Zeger SL, editors, Springer, New York.

Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese, JC, Schroeder M, Brown PO, Botstein D, Sherlock D (2003). The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res*. 31: 94-6.

Irizarry RA, Hobbs B, Colin F, Beazer-Barclay YD, Antonellis K, Scherf U, Speed TP (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, 4: 249-64.

Kamberova G, Shah S (2002). *DNA Array Image Analysis: Nuts & Bolts*, DNA Press.

Kim JH, Shin DM, Lee YS (2002). Effect of local background intensities in the normalization of cDNA microarray data with a skewed expression profiles. *Exp Mol Med*. 34: 224-32.

Nunez-Garcia J, Mersinias V, Cho KH, Smith CP, Wolkenhauer O (2004). A study of the statistical distribution of the intensity of pixels within spots of DNA microarrays: What is the appropriate single-valued representative?, *Appl Bioinformatics*, 2: 229-239.

Quackenbush J (2002). Microarray data normalization and transformation, *Nature Genetics*, 32: 496-501.

Pat Brown Laboratory Protocol (2002) (Online) Department of Biochemistry, Stanford University, http://cmgm.stanford.edu/pbrown/protocols/index.html.

Schena M (2000). *Microarray Biochip Technology*, Eaton.

Stekel D (2003). *Microarray Bioinformatics*, Cambridge University Press.

Suite M (1999). *User Guide for Use with IPLAB for Macintosh*, Scanalytics Inc., Fairfax, VA. Supplement.

Wit E, McClure J (2004). *Statistics for Microarrays.* John Wiley & Sons.

Xiao Y, Hunt CA, Segal MR, Yang YH (2004). A Novel Stepwise Normalization Method for Two-Channel cDNA Microarray. *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, San Francisco, CA, USA. pp. 2921-4.

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

Yang YH, Dudoit S, Luu P, Speed TP (2001). Normalization for cDNA microarray data, In: *Microarrays: Optical Technologies and Informatics.* Bittner ML, Chen Y, Dorsel AN, Dougherty ER (eds.), , Volume 4266 of Proceedings of SPIE.

Yang YH, Speed T (2003). Design and analysis of comparative microarray experiment. In *"Statistical Analysis of Gene Expression Microarray Data"*. Chapman & Hall/CRC Press, pp: 35-92.