Short Technical Report

Local Mean Normalization of Microarray Element Signal Intensities across an Array Surface: Quality Control and Correction of Spatially Systematic Artifacts

BioTechniques 32:1316-1320 (June 2002)

Carlo Colantuoni[1,2], George Henry[1], Scott Zeger[3], and Jonathan Pevsner[1,2]

[1]Kennedy Krieger Institute, [2]Johns Hopkins University School of Medicine, and [3]Johns Hopkins School of Public Health, Baltimore, MD, USA

ABSTRACT

Here we present a methodology for the normalization of element signal intensities to a mean intensity calculated locally across the surface of a DNA microarray. These methods allow the detection and/or correction of spatially systematic artifacts in microarray data. These include artifacts that can be introduced during the robotic printing, hybridization, washing, or imaging of microarrays. Using array element signal intensities alone, this local mean normalization process can correct for such artifacts because they vary across the surface of the array. The local mean normalization can be used for quality control and data correction purposes in the analysis of microarray data. These algorithms assume that array elements are not spatially ordered with regard to sequence or biological function and require that this spatial mapping is identical between the two sets of intensities to be compared. The tool described in this report was developed in the R statistical language and is freely available on the Internet as part of a larger gene expression analysis package. This Web implementation is interactive and user-friendly and allows the easy use of the local mean normalization tool described here, without programming expertise or downloading of additional software.

INTRODUCTION

Experimental procedures in the creation and use of DNA microarrays often introduce artifacts into the resulting data. These artifacts are often spatially systematic, varying across the surface of the microarray. Such experimentally introduced variability in element signal intensities can have profound effects on gene expression values and can result in the misidentification of differentially expressed genes. Terrence Speed and collaborators have noted these effects resulting from the robotic printing of microarrays and have proposed methods for their correction, based on the normalization of array elements derived from the same robotic arrayer print tip (http://www.stat.Berkeley.EDU/users/terry/zarray/Html/papersindex.html). Here we present a more general methodology for the correction of artifacts that can be identified across the physical surface of microarrays. This entails the use of x and y coordinates for each array element in the normalization of each element intensity to a mean intensity that is calculated locally across the surface of the microarray.

These methods endeavor to detect and correct any artifact that introduces noise into the microarray data systematically across the surface of the array. These include artifacts such as those described by Speed et al. that result from the robotic printing process, hybridization artifacts including nonuniform background signal intensities, and any other spatially systematic artifact that is introduced during the production and use of microarrays. To illustrate the utility of these local mean normalization methods, here we apply them to an example dataset, correcting for nonuniform background intensities that result from radioactive hybridization methods.

The algorithms described here were developed in the freely available R statistical language (http://www.r-project.org). To make this tool available to biologists who lack R programming expertise, we created an interactive, user-friendly Web implementation of this tool that is available at http://pevsnerlab.kennedykrieger.org/snomad.htm. No programming expertise or download of additional software is required for researchers seeking to apply this tool to their gene expression data.

MATERIALS AND METHODS

Global Mean Normalization

Global normalization methods should be carried out on raw array element intensities before the application of the local normalization. While global mean normalization is rendered computationally obsolete by the local mean normalization detailed below, it is very

useful in the intermediate steps of analysis and in data visualization. For this reason, we recommend applying the global mean normalization before the local mean normalization (i.e., division of each individual raw element intensity by the global mean intensity).

Several sources of variance may have a constant impact on all element signal intensities in a dataset, such as differential label incorporation into probes, differing amounts of probes used, or differences in detection efficiency. To correct for this uniform variance, global normalization processes are applied to nearly all DNA microarray data. Both one- and two-channel array datasets require global normalization before the analysis of expression values or ratios. In one-channel datasets, this entails the division of each element signal intensity by a correction factor. Two-channel datasets are best handled as two individual one-channel datasets in this respect, such that each element signal intensity is divided by a correction factor that is specific to the channel from which it was derived.

This correction factor can be calculated in a number of different ways. Perhaps the most common is the normalization of all element signal intensities to the mean intensity of all elements contained within a single array (for one-channel data) or single channel (for two-channel data):

$$\text{Element Intensity} / \text{Mean Element Intensity} = \text{Global Mean Normalized Intensity} \qquad \text{[Eq. 1]}$$

The global mean normalization ensures that the mean gene expression value or ratio for individual datasets will be equal to a value of 1.0 and equivalent across datasets. Only elements intended to measure gene expression should be involved in this normalization. Controls, total genomic DNA, and other similar elements should be excluded. When using this method of normalization, it is best that the number of genes involved be large (several thousand) and that the samples to be compared are highly similar (e.g., from the same tissue source). Similarly, it is inadvisable to construct an array consisting entirely of genes that are expected to change or are all thought to be involved in the process being studied. The design of such arrays complicates the normalization of the resulting gene expression values.

Normalization to a subset of genes designated as invariant, or "housekeeping genes", is also a common practice. This can lead to problems if any of the genes contained within this subset are regulated in the experimental paradigm under study. Without prior validation of gene subsets for the particular samples being studied, it is not easy to determine which gene expression will be invariant. Another alternative is to add known amounts of labeled cDNAs not otherwise present in the samples being studied, but included on the array. Such additions require the precise quantification of RNA before the addition and should be made at the earliest possible point in the sample preparation. These intensity values can then be used for global normalization.

Local Mean Normalization: Correction of Spatially Variable Hybridization Artifacts

As an example of how the local mean normalization can be used to detect and correct artifacts that vary across the surface of a microarray, we applied these methods to an example dataset. Significant, nonuniform background intensities resulting from radioactive hybridization methods are present in this dataset (possibly due to uneven hybridization conditions, membrane drying, or uneven washing). Using element signal intensities alone, without any measurement of background intensity, the local mean normalization across the surface of the microarray can correct for this uneven background signal. Figure 1A depicts the comparison of two independent one-channel gene expression datasets derived from a radioactivity-based microarray system (Human GeneFilters®; Research Genetics, Huntsville, AL, USA). The mRNA derived from two human postmortem brain tissue samples was used to generate radiolabeled cDNA probes for hybridization to the approximately 5500 cDNAs spotted on a nylon filter. Hybridized arrays were imaged using standard phosphorimaging methods. The Pathways™ software from Research Genetics was used to ex-

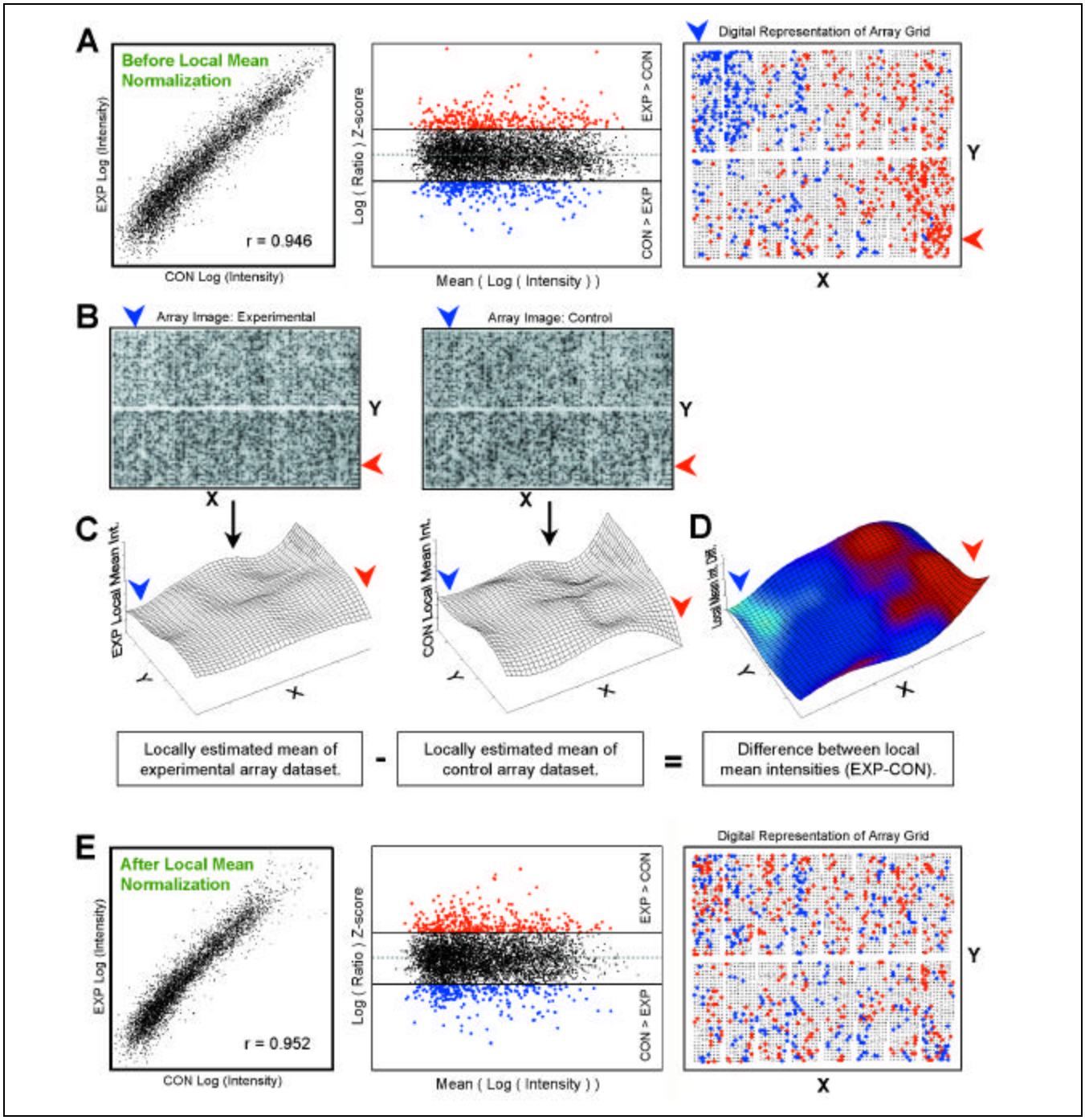

Figure 1. Local normalization across microarray surface for the correction of variable background intensity. The data depicted were generated using mRNA derived from postmortem human brain tissue that was analyzed using the Human GeneFilter microarray system. The visualization and analysis applied to the gene expression values in panels A and B are detailed in more depth in the text. (A) Comparison of gene expression values from two independent one-channel datasets before the local array surface fit correction. The left panel depicts a scatter plot of the $\log_{10}$ intensities for the two samples. The center panel shows an intensity-versus-ratio scatter plot. Array elements identified as up- and down-regulated (center panel and red and blue points, respectively) localize to particular regions of the array surface (right panel and red and blue arrowheads). Array elements highlighted in the intensity-versus-ratio scatter plot (center panel) correspond to those highlighted in the representation of the array grid surface (right panel). (B) The regions to which these genes localize correspond to regions of the array surface that demonstrate especially different background intensities between the two arrays (both panels, red and blue arrowheads). (C) A robust local regression (loess in the R statistical language) is used to estimate the mean element signal intensity locally across the 2-D surface of each array. (D) The difference between these local mean estimates highlights the regions of the arrays that will undergo the heaviest normalization (red and blue arrowheads). (E) Comparison of gene expression values after local mean normalization. The amount of correlation between the two datasets has increased (compare the Pearson's correlation coefficient, r, in panels A and E). Array elements highlighted in the intensity-versus-ratio scatter plot (center panel) correspond to those highlighted in the representation of the array grid surface (right panel). Array elements identified as differentially expressed (center panel) no longer show systematic distribution across the array surface (right panel). CON, control, and EXP, experimental.

tract quantitative information from the resulting images.

Our basic microarray data analysis began with a scatter plot of Log_{10} intensities (Figure 1, A and E, left panels), finally resulting in an intensity-versus-ratio scatter plot (Figure 1, A and E, center panels). This plot displays each gene (i.e., microarray element) as an expression ratio (y-axis) at a particular expression level (x-axis). Therefore, while the x-axis value depicts the average level at which a gene is expressed, the y-axis value is reflective of the differential expression between the two samples under study. Differential gene expression values (y-axis) are expressed in standard deviation units derived from Log_{10} (ratio) calculations (Z-scores). At this point, it is appropriate to apply a uniform differential expression cut-off across all of the levels of expression in the dataset (Figure 1, A and E, center panel, horizontal value reference lines). To illustrate the utility of this method of local mean normalization, the entire analysis is carried out on an example dataset before (Figure 1A) and after (Figure 1E) the application of the local mean normalization detailed below. The computational tools involved in this basic analysis are available along with the local mean normalization tool detailed in the SNOMAD Web site (http://-pevsnerlab.kennedykrieger.org/snomad.htm). Figure 1, B–D, details the local mean normalization process.

Figure 1A, left panel, depicts a scatter plot of globally normalized log element intensities from each array. Each gene is plotted as a single point, with an intensity derived from each of the two microarray experiments. Figure 1A, center panel, highlights the array elements that are identified as differentially expressed in the first analysis. When the data are represented as points in their original physical position (Figure 1A, right panel), it is clear that these array elements localize to particular regions of the array surface (corresponding highlighted points and red and blue arrowheads). The regions to which these genes localize correspond to regions of the array surface where the difference in background signal intensity between the two arrays is especially high (Figure 1B, both panels, red and blue arrowheads). Thus, the majority of the array elements identified in this analysis would be misidentified as being differentially expressed between the two samples under study.

To correct for this artifact, we propose the normalization of all array element signal intensities to a mean intensity that is estimated locally across the 2-D surface of each microarray. This mean intensity is estimated using the "loess" function in the R statistical language. The loess function or locally weighted regression (2,3) is a method to estimate the conditional mean of one variable (element signal intensity) as a function of a second variable (position on the microarray surface). Therefore, the loess function is used to calculate the mean element signal intensity at each point across the array surface using intensities at neighboring points. The smoothness of the loess curve (i.e., the size of the window used in the calculation of this local mean intensity) is controlled by a bandwidth parameter, called the span, which can be determined by the user. This is a robust procedure in that the mean intensity estimate is insensitive to a fraction of outlying or extreme intensities. This fraction, called the trim, can also be determined by the user.

Figure 1C illustrates the loess fitting procedure, in which the locally estimated mean intensity is plotted as a smooth function across the spatial surface of each array. Because the global mean normalized intensities were used as the input into the analysis, these fits can be visualized on identical scales (Figure 1C) and subtracted from one another to identify differences in the locally estimated mean intensity between the two arrays (Figure 1D). This difference plot is also provided in Web implementation of this tool and can be used as a method to detect artifacts that are systematically distributed across the array surfaces. The greatest differences in this locally estimated mean intensity (Figure 1D, arrowheads) correspond to the same regions of the array surface to which, apparently, differentially regulated genes localize (Figure 1A, right panel, arrowheads) and background intensities vary most between the two array experiments (Figure 1B, highlighted points). This again indicates that the genes identified in this first analysis are not differentially expressed between the two samples

but simply happen to lie in regions of the array surface where hybridization artifacts are especially intense and different between the two arrays.

To ensure that the differences in this locally estimated mean (Figure 1D) reflect unwanted hybridization artifacts in the expression data, two conditions of the array design must be satisfied. (*i*) The arrays being compared in this way must share identical spot mapping. That is, the content and position of all of the array elements must be identical between the two arrays being compared. (*ii*) The spot mapping must be random from a biological point of view. That is, the genes with related functions or sequences should not be spatially related on the array. For most large microarrays (e.g., >10 000 elements), this is not usually a problem. This issue is more important in the analysis of data derived from commercial or custom-designed arrays that are focused on particular biological questions.

In using element signal intensities to correct for variability in background signal intensity, it is imperative that the element intensity from each corresponding location on the two arrays be derived from an identical array element. The first of the previously described conditions can ensure this. The differential expression of a group of genes that are spatially related on the array surface could also disrupt this analysis. The second condition ensures that biologically related genes are not spatially related on the array surface so that this occurrence is unlikely.

Using the locally estimated mean intensities (Figure 1C), we can now normalize all individual element intensities:

Corrected Element Intensity = Element Intensity/Local Mean Intensity [Eq. 2]

This is similar to a global mean normalization:

Corrected Element Intensity = Element Intensity/Global Mean Intensity [Eq. 3]

However, Global Mean Intensity is a single value used to normalize element intensities at all array positions, while Local Mean Intensity is a smooth function, estimated locally across the array surface and used to normalize array elements at corresponding positions on the array surface.

Element intensities from each array experiment are plotted against each another before (Figure 1A) and after (Figure 1E) the local mean normalization is applied to each of the datasets. The Pearson's product moment correlation coefficient, r, is slightly greater after the local surface normalization (compare Figure 1, A and E, left panels). Array elements identified as differentially expressed (Figure 1E, center panel) no longer show the systematic distribution across the array surface (Figure 1E, right panel).

DISCUSSION

On the Normalization of Microarray Data

Several groups have addressed basic normalization processes such as background subtraction and global mean normalization (1,4–7). Sources of variance in gene expression datasets that are spatially systematic across the surface of the microarray are not rare. Spatially systematic artifacts can be corrected for in microarray data using the spatially driven local normalization processes detailed here. Background correction is just one of the many practical uses of local normalizations in the refinement of gene expression data. We have previously described additional local normalization processes that are local across different dimensions in microarray data (see the SNOMAD Web site).

Public Use of These Algorithms

The freely available R statistical language (http://www.r-project.org) was used to develop the local mean normalization tool detailed in this report. To make this tool available to researchers who lack expertise in this programming language, an HTML form in combination with PERL scripts were used to create an interactive Web implementation of this tool that is available at http://pevsnerlab.kennedykrieger.org/snomad.htm. Internet access and a standard HTML browser are the only system requirements necessary for the full use of this tool.

Using this Web page, users can upload their own gene expression data as a tab-delimited text file. PERL scripts then assemble and execute the appropriate R code. The results of the request are then returned via a new HTML page or as an e-mail attachment. The results include a text file containing numeric values and image files depicting many of the graphical representations included in Figure 1.

ACKNOWLEDGMENTS

Funding for this work was provided by MRRC grant nos. HD24061-12 to J.P. and NIH RO1 MH56639 to S.Z.

REFERENCES

1. **Beissbarth, T., K. Fellenberg, B. Brors, R. Arribas-Prat, J. Boer, N.C. Hauser, M. Scheideler, J.D. Hoheisel, et al.** 2000. Processing and quality control of DNA array hybridization data. Bioinformatics *16*:1014-1022.
2. **Cleveland, W.S.** 1981. LOWESS: program for smoothing scatterplots by robust locally weighted regression. Am. Stat. *35*:54.
3. **Hastie, T. and R. Tibshirani.** 1990. Generalized Additive Models, 1st ed. Chapman and Hall, London.
4. **Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J.E. Hughes, E. Snesrud, et al.** 2000. A concise guide to cDNA microarray analysis. BioTechniques *29*:548-556.
5. **Liao, B., W. Hale, C.B. Epstein, R.A. Butow, and H.R. Garner.** 2000. MAD: a suite of tools for microarray data management and processing. Bioinformatics *16*:946-947.
6. **Schuchhardt, J., D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, and H. Herzel.** 2000. Normalization strategies for cDNA microarrays. Nucleic Acids Res. *28*:E47.
7. **Smid-Koopman, E., L.J. Blok, S. Chadha-Ajwani, T.J. Helmerhorst, A.O. Brinkmann, and F.J. Huikeshoven.** 2000. Gene expression profiles of human endometrial cancer samples using a cDNA-expression array technique: assessment of an analysis method. Br. J. Cancer *83*:246-251.

Received 18 June 2001; accepted 8 March 2002.

Address correspondence to:

Dr. Jonathan Pevsner
Dept. of Neurology
Kennedy Krieger Institute
707 N. Broadway
Baltimore, MD 21205, USA
e-mail: pevsner@kennedykrieger.org