Topographical imaging technique for qualitative analysis of microarray data

James Ch. Davis, Anna-Lisa Paul, Robert J. Ferl, and Mark W. Meisel
University of Florida, Gainesville, FL, USA

BioTechniques 41:554-558 (November 2006)
doi 10. 2144/000112288

While high-density microarrays present the means of monitoring genome-wide patterns of gene expression, the question of how to analyze or even present the data for tens of thousands of genes remains a challenge. One method to graph changes in gene expression between two samples involves a two-dimensional (2-D) plot, also referred to as a scatterplot, generated by assigning each gene a set of coordinates corresponding to the measured expression in each sample. These plots obscure useful information, however, when a large gene count causes saturation (i.e., complete coverage) across areas of the graph (Figure 1). Although additional diagnostic analyses can be performed (1), the additional sets of plots and numerical results are not immediately transparent. Alternatively, topographical plots (i.e., relief maps showing the third dimension), which have widespread interdisciplinary applications, can be generated (2). In fact, freeware routines (3,4) and commercial imaging/analysis packages (5,6) are examples of available tools.

Here we describe a simple MATLAB® program (www.mathworks.com), UFGenie (available at www.phys.ufl.edu/~meisel/bioinformatics.html) that generates topographical plots of microarray data sets. This routine is useful for quick qualitative analysis that reveals significant information, especially with respect to regions of low differential gene expression not visible in standard analysis. Once anomalous trends are qualitatively identified, the data can be quantitatively analyzed and mined to identify the source of the variations. This paper provides a brief overview of our approach, and a comparison of our method to other tools is provided in the supplementary material (available online at www.BioTechniques.com).

When 2-D gene response graphs are generated, a logarithmic scale can be used to show the more biologically significant fold-change for each gene (Figure 1). In this type of plot, the points along the diagonal represent genes that exhibited identical expression between the two treatments. In addition, data falling within the 2-fold up or down differential expression, designated by the off-diagonal lines in Figure 1, are often considered to represent insignificant changes. This generic cutoff for identifying significant change can be refined by biological replication and statistical analysis of variance. The data chosen for an axis in this type of graph could represent a single array or a large set of biological replicates that have been averaged and normalized. While comprehensive and visually appealing, Figure 1 demonstrates the common situation in which a large gene count causes saturation over a significant portion of the graph. This saturation could cause a biological response to be hidden or could mask meaningful variation among samples. However, by initially binning the data from two arrays, topographical plots can be created where the third (color) dimension represents the number of genes present at any given expression level (Figure 2). While previous analyses focused only on genes with changes in expression greater than some variance-based limit (often taken as 2-fold), the topographical plots yield gene density information both inside and outside this region.

The data used for Figures 1 and 2 were derived from the ATH1 GeneChip® arrays (Affymetrix, Santa Clara, CA, USA) measuring gene expression in *Arabidopsis* leaf tissue subjected to different environmental stresses (subjection to varying pressures: 25 kPa for Array A, 101 kPa for Array B, and 75 kPa for array C). Although Affymetrix chips were used in these example data sets, the process is easily extended to other arrays and is independent of any particular normalization or smoothing protocol. Since multiplicative changes are more indicative of substantial gene response, the base ten logarithms of probe spot intensity data were calcu-

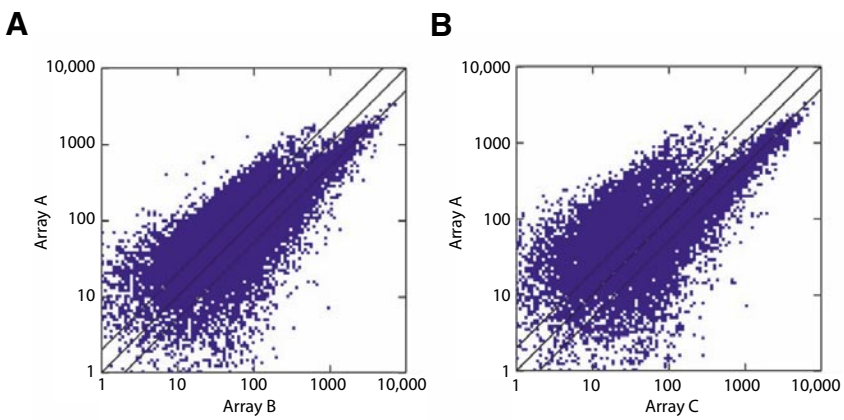

Figure 1. Saturated scatterplots. Two saturated scatterplots where the axes are logarithmic. The center black line indicates the area of zero expression change between treatments, while the other two parallel lines represent the 2-fold expression increase or decrease regions. Each individual gene is depicted by a single data point whose coordinates correspond to the measured differential expression on each sample. Large groups of data both inside and outside the 2-fold differential response region are obscured because of high gene density. Panels A and B appear similar in this two-dimensional representation, and variations in the high-density regions are invisible. Data for these arrays were obtained from gene expression in leaf tissues of *Arabidopsis* plants using ATH1 GeneChip arrays.

lated before being binned and plotted. The large number of genes present (up to approximately 22,000) permits the data to be binned into regions substantially smaller than experimental uncertainty, while large enough to contain substantial gene groups. Furthermore, a small group of bins may often contain a large number of genes relative to other graph areas. Consequently, a threshold cutoff value is required to visualize the low-level topographical texture of the graph. Both the cutoff value and bin size are user-adjustable.

The profiles examined in Figure 2 show unexpected shifts between sample pairs, which are not apparent in the corresponding 2-D plots (Figure 1). More specifically, the topographical plots reveal a bifurcation of the high-density region in Figure 2A, while the data in Figure 2B exhibit a weaker effect. Consequently, a new level of interpretation is available through quick visual inspection, and the decision to perform quantitative analysis can be more selective, thereby increasing the efficiency of the scanning process.

A principle advantage of the topographical plots is the ability to detect variations at low differential gene expression levels. Differential expression <2-fold is not considered significant, and data are not normally mined in this region, even though this range is often saturated with genes. Furthermore, despite an uncertainty in the response of individual genes, the collective density and spread within this 2-fold region may vary from treatment to treatment in a significant way. In other words, changes in the characteristic gene response spread within this area are a reliable indication of biological or technical variation, even though the differences may not be apparent elsewhere. Our analysis technique affords the opportunity to observe these variations, as shown in Figure 2, where the intense response of each plot is shifted and does not center about the diagonal.

Furthermore, the topographical plots have application in evaluating the fidelity of biological replications on a genome-wide scale. For example, in a plot comparing two biological replicates—if every effort was made to prepare identical tissues—the data would ideally be confined entirely to the diagonal. However, because the variance in some expression may be large in equivalent environments (7), the total data spread (χ^2 for example) is not necessarily the best indicator of biological equivalent control samples. Alternatively, the area of the intense response region about the diagonal is a way to assess biological similarity and measures the ability to produce accurate biological replicates without considering outlying genes. In other words, the intense response area of topographical plots reflects sample similarity, even in the absence of notable changes outside the 2-fold region.

A final application of our technique relates to the question of gene chip reliability. Variance in microarray data are reduced when data from multiple chips are averaged and the resulting mean is considered. This averaging procedure can include elimination of specific genes if variance from chip to chip is considered inappropriately high or the values are otherwise indicative of error (8). Topographical plots of nonaveraged expression data potentially reveal either a compromise or an unusual positive response in an entire single array, and the bifurcation in Figure 2 may be an example of these possibilities that would not be as readily observable in averaged data. Furthermore, the intense off-diagonal response that is within the 2-fold up/down response region is also information that is not available in a 2-D plot.

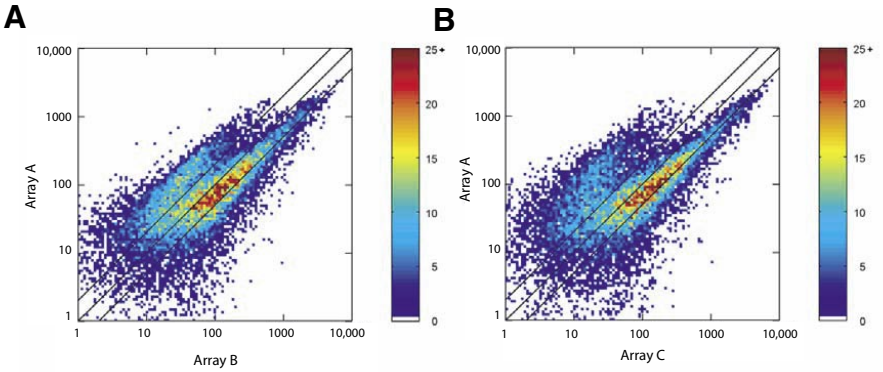

Figure 2. Topographical plots. Topographical plots of the same data in Figure 1 show a significant shift between panels A and B in previously saturated areas. The number of genes in each bin is designated by a linear (color) scale as indicated by the bar. A strongly pronounced bifurcation of data are seen in panel A, while panel B exhibits a smoother distribution. This change could be biologically motivated or indicative of a systematic error. Regardless of the source, conventional analysis is likely to miss these trends.

ACKNOWLEDGMENTS

This work was supported, in part, by the National Science Foundation (NSF) through DMR-0305371, NASA grant no. NNA04CC61, and by the University of Florida University Scholars Program. We acknowledge enlightening conversations with S.J. Hagen and M. Popp.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

1. **Park, T., S.-G. Yi, S.Y. Lee, and J.K. Lee.** 2005. Diagnostic plots for detecting outlying slides in a cDNA microarray experiment. BioTechniques *38*:463-471.
2. **Carr, D.B., R.J. Littlefield, W.L. Nicholson, and J.S. Littlefield.** 1987. Scatterplot matrix techniques for large n. J. Am. Stat. Assoc. *82*:424-436.
3. **Carr, D.B., N. Lewin-Koh, and M. Maechler.** 2005. hexbin: hexagonal binning routines. R package version 1.4.0.
4. **Eilers, P.H.C. and J.J. Goeman.** 2004. Enhancing scatterplots with smoothed densities. Bioinformatics *20*:623-628.
5. **Dysvik, B. and I. Jonassen.** 2001. J-Express: exploring gene expression data using Java. Bioinformatics *17*:369-370.

6. **Werner-Washburne, M., B. Wylie, K. Boyack, E. Fuge, J. Galbraith, J. Weber, and G. Davidson.** 2002. Comparative analysis of multiple genome-scale data sets. Genome Res. *12*:1564-1573.
7. **Zakharkin, S.O., K. Kim, T. Mehta, L. Chen, S. Barnes, K.E. Scheirer, R.S. Parrish, D.B. Allison, and G.P. Page.** 2005. Sources of variation in Affymetrix microarray experiments. BMC Bioinformatics *6*:214.
8. **Iacobas, A.D., M. Urban, A. Massimi, and D.C. Spray.** 2002. Improved procedures to mine data obtained from spotted cDNA arrays. J. Biomol. Tech. *13*:5-19.

Received 7 July 2006; accepted 22 August 2006.

Address correspondence to Mark W. Meisel, Department of Physics, University of Florida, Gainesville, FL 32611, USA. e-mail: meisel@phys.ufl.edu