SUPPLEMENTARY MATERIAL FOR:

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)

Comparison of UFGenie to Similar Programs

Currently, very few methods are available to easily generate scatterplots that do not saturate with high gene numbers, effectively obscuring useful information. UFGenie solves this problem through data binning and projection into three-dimensions (3-D). This supplementary report provides a comparison of our solution against other protocols that attempt to solve the problem of data point saturation in two-array scatterplots.

Since most programs that generate similar graphs as UFGenie rely heavily on smoothing techniques, understanding the benefits and costs of such procedures is important. Large data sets lend themselves well to binning without additional smoothing protocols. Supplementary Figure S1A contains approximately 17,000 genes and thus standard binning reflects data trends clearly and accurately. A smoothed version (1) of the same data, shown in Supplementary Figure S1B, shows extrapolation into empty regions, and even loses some outlying genes. Additionally, smoothed data requires a color scale that approaches zero gradually and tends to obscure single data points (notice the difference in color scales between Figure S1, A and B). A nonsmoothing routine preserves a truer representation of the data and reveals general response patterns when many thousands of genes are plotted. This digitization of the field also allows for a clearer color distinction between 0 and 1 count, thereby giving the color scale more concrete meaning.

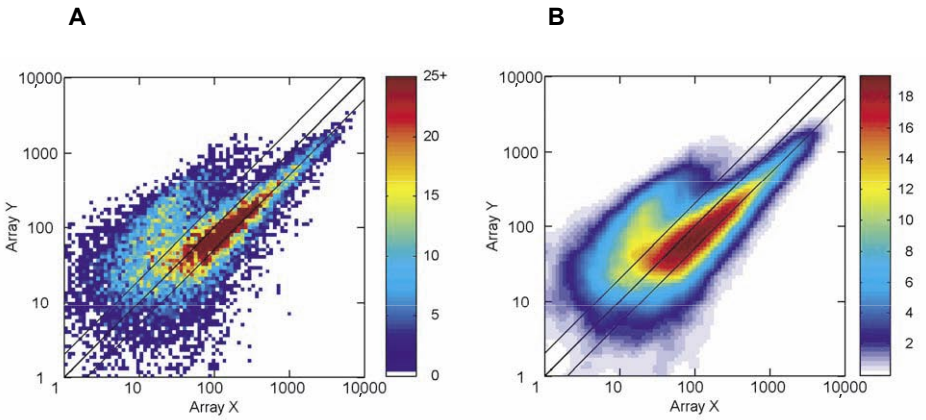
In Supplementary Figure S2, however, the total gene number is severely reduced (down to approximately 7000). As Figure S2B shows, the same data can be smoothed with a simple algorithm, as reported by Eilers and Goeman in (1), to make the trend more appealing in some respects. This optional smoothing protocol respects boundaries, meaning that it does not assume zero counts past the boundaries of the scatterplot. Differences between this smoothing procedure and a weighted-mean protocol become a factor when data approach the experimental boundaries. This problem is especially relevant to gene expression data, since scatterplots commonly have a bulk of data around zero. Supplementary Figure S3 demonstrates the major discrepancy between the smoothing routine outlined in (1) and a weighted-mean distribution generated in J-Express™ on the same data (2). The density falls off dramatically toward the boundaries in Figure S3B, while Figure S3A shows no artificial density decrease around the graph edges. In Figure S3B, the gene density is reduced because the program tacitly assumes a density of zero in outlying graph regions.

Large database mining programs such as VxInsight™ (3) provide 3-D visualization of data sets where a third dimension represents point density. Our protocol easily duplicates this 3-D depiction of data, as topographical plots can be naturally extended into three dimensions. Through the simple command prompt interface, users may opt to produce two-dimensional (2-D) topographic plots or full 3-D surfaces. The surfaces allow for full rotation and inspection from any angle, both above and below. Mapping the color scale onto a third spatial dimension in this manner often improves clarity, since a subjective color scale makes 1:1 density comparisons more difficult. The surface height, however, scales with gene density in an obviously linear fashion. Supplementary Figure S4 shows the same two arrays presented in Supplementary Figure S1, smoothed and projected into 3-D.

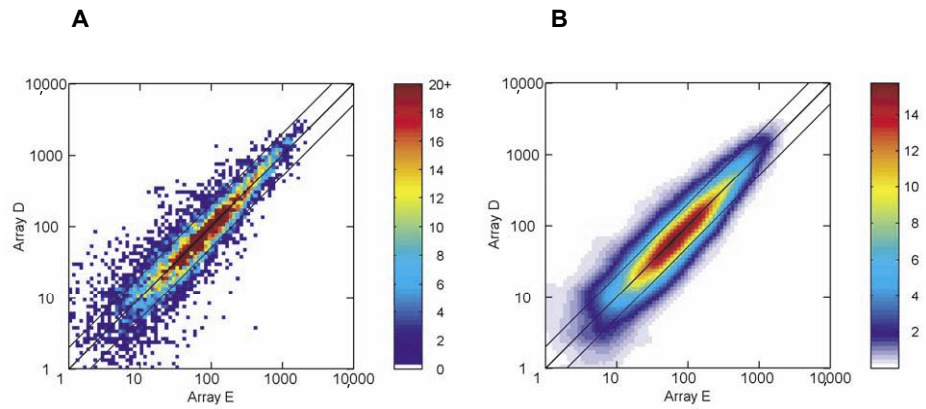
Finally, the R Hexbin package (4) employs some of the same concepts as UFGenie, but it lacks several features integrated into our program. Specifically, UFGenie is specially designed for use in microarray analysis, and as such, it automatically implements optional log-scale plotting, a user-adjustable maximum height cut-off height, and text prompts for a shallow learning curve. Additionally, UFGenie allows gene data to be input and plotted with far fewer commands; data can simply be copied and pasted from a Microsoft® Excel® spreadsheet or ASCII text file. Supplementary Figure S5 shows the same gene data sets binned with UFGenie (Figure S5A) and using Hexbin with a similar color scale (Figure S5B). Figure S5A allows for greater visibility of the low-level topographical texture because bin heights above 20 share the same color.

REFERENCES

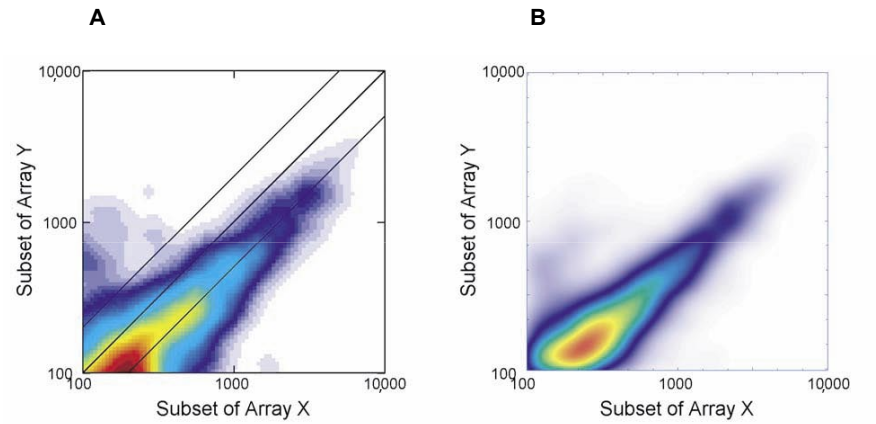
1. **Eilers, P.H.C. and J.J. Goeman.** 2004. Enhancing scatterplots with smoothed densities. Bioinformatics *20*:623-628.
2. **Dysvik, B. and I. Jonassen.** 2001. J-Express: exploring gene expression data using Java. Bioinformatics *17*:369-370.
3. **Werner-Washburne, M., B. Wylie, K. Boyack, E. Fuge, J. Galbraith, J. Weber, and G. Davidson.** 2002. Comparative analysis of multiple genome-scale data sets. Genome Res. *12*:1564-1573.
4. **Carr, D.B., N. Lewin-Koh, and M. Maechler.** 2005. hexbin: hexagonal binning routines. R package version 1.4.0.



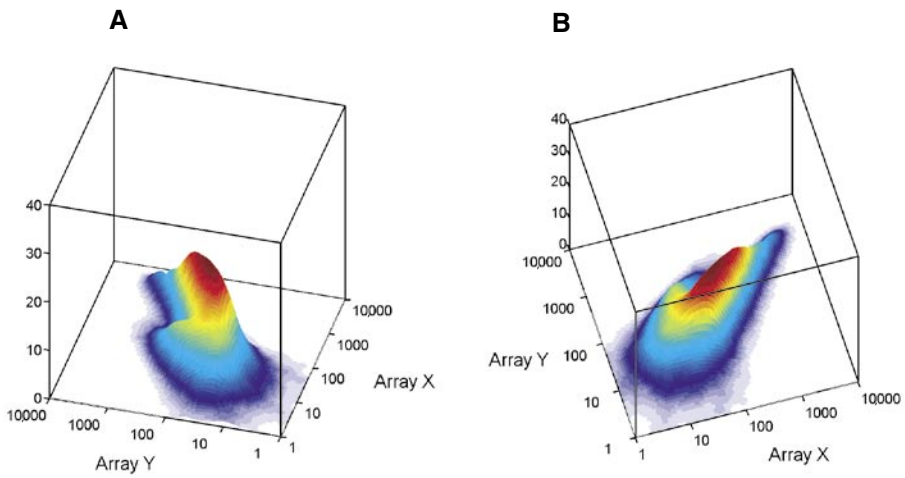
Supplementary Figure S1. Smoothing on a large data set. Panel A shows measured expression of two gene arrays where each gene is assigned an X and Y value corresponding to its expression in two respective arrays. The XY scales are logarithmic, and the third dimension (color scale) is linear. The perfect diagonal marks no change in expression between the two arrays, while the off-diagonal lines mark the 2-fold up- and down-regulation bounds. Once binned, each bin is assigned a color based on the number of genes it contains, as determined by the color bar. There are multiple smoothing protocols that can make array data more visually appealing. The same arrays are plotted in panels A and B, except UFGenie's integrated smoothing routine (as described by Eilers and Goeman in Reference 1) was run on panel B.



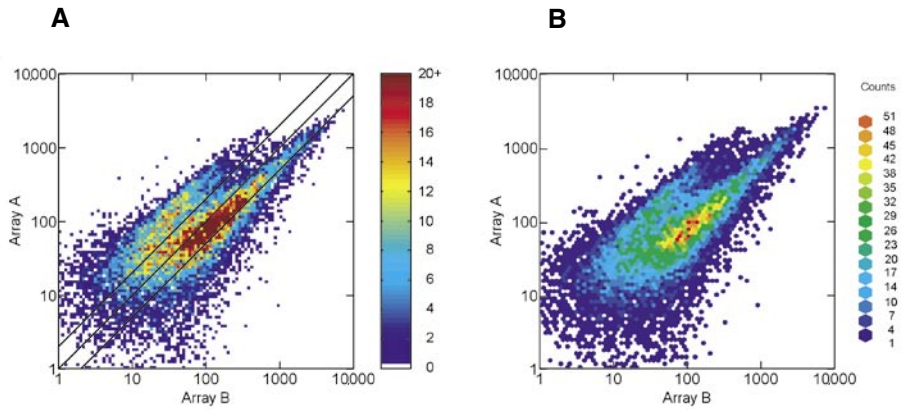
Supplementary Figure S2. Smoothing on a smaller data set. Measured expression in two gene arrays plotted against one another. Again, the diagonal marks no change in expression between the two arrays, and the off-diagonal lines mark the 2-fold up- and down-regulation bounds. Although outliers are masked by the smoothing procedure, panel B maintains the trends apparent in the digitized plot of panel A. In the case of smaller data sets, some data points are hidden more easily by smoothing, but smoothing may also produce a more comprehensible trend.



Supplementary Figure S3. Two smoothing algorithms with mixed results. The same subset of data plotted in Supplementary Figure S1 is shown here, as smoothed by (A) UFGenie based on the algorithm in Reference 1 and (B) J-Express. The high-density (red) region in panel B cannot extend to the edges of graph because weighted-mean smoothing algorithms assume zero genes past the graph bounds. The axes in panel B were labeled outside J-Express to improve clarity.



Supplementary Figure S4. Three-dimensional (3-D) plots. The same data set presented in Supplementary Figure S1 is here extended into 3-D. Partial rotations can present an intuitive picture of data distribution without having to rely on a color scale alone. Additionally, the MATLAB® figures can be smoothly turned and viewed from any angle, both from above and below the surface.



Supplementary Figure S5. Hexbin generated graphic. The same data as presented in Supplementary Figure S2 from the main article is binned in (A) with UFGenie and in (B) with the R hexbin package (4). Unexpected data bifurcations are revealed in both panels. The plot made with Hexbin does not limit the maximum bin count, making the low-level topographic texture of the map less apparent. The axes of panel B were edited outside of R hexbin to improve visibility.