

## Short Technical Report

# MarC-V: A Spreadsheet-Based Tool for Analysis, Normalization, and Visualization of Single cDNA Microarray Experiments

BioTechniques 32:338-344 (February 2002)

**J.J. Schageman, M. Basit,  
T.D. Gallardo, H.R. Garner,  
and R.V. Shoheit**

The University of Texas  
Southwestern Medical  
Center, Dallas, TX, USA

### ABSTRACT

*The comprehensive analysis and visualization of data extracted from cDNA microarrays can be a time-consuming and error-prone process that becomes increasingly tedious with increased number of gene elements on a particular microarray. With the increasingly large number of gene elements on today's microarrays, analysis tools must be developed to meet this challenge. Here, we present MarC-V, a Microsoft® Excel® spreadsheet tool with Visual Basic® macros to automate much of the visualization and calculation involved in the analysis process while providing the familiarity and flexibility of Excel. Automated features of this tool include (i) lower-bound thresholding, (ii) data normalization, (iii) generation of ratio frequency distribution plots, (iv) generation of scatter plots color-coded by expression level, (v) ratio scoring based on intensity measurements, (vi) filtering of data based on expression level or specific gene interests, and (vii) exporting data for subsequent multi-array analysis. MarC-V also has an importing function included for GenePix® results (GPR) raw data files.*

### INTRODUCTION

With spotted microarrays (cDNA and oligonucleotide) becoming the most (7) prominent and powerful meth-

od for studying global gene expression levels and expression profiling, data management and analysis tools must evolve to accommodate the extremely large datasets resulting from these experiments. These data are generated from the scanning of a microarray slide with a two-laser scanner to acquire intensity measurements for each spotted array element. In a two-color, or two-channel, experiment, two fluorescently labeled (Cy3 and Cy5) RNA probes are mixed and hybridized to PCR amplified mRNA transcripts on a glass slide. The "wet-lab" steps in this process (as well as further background and optional analysis) are well described by Hegde et al. (5). The extraction of intensity measurements from both channels typically yields microarray data that are in a tab-delimited text format spanning multiple columns and thousands of rows, depending on the density of microarray elements. Manual cutting, pasting, plotting, and calculation of aggregate statistics from these data are error-prone and often daunting tasks.

Most microarray scanners typically come with analysis software to accommodate the need to sort and visualize vast datasets. One common limitation to these software packages is the difficulty in customizing gene expression analysis to laboratory specifications. Examples of such customizations include allowing the user to apply lower-bounds thresholding or flagging specific gene interests such that the associated differential expression ratios may easily be viewed without sorting through the entire dataset. Many groups using microarray data solve some of these problems using popular spreadsheet programs such as

Microsoft® Excel® that contain customizable formulas for accomplishing the complex calculations required for microarray data analysis. Here, we report on an enhanced version of an Excel 2000 workbook containing Microsoft Visual Basic for Applications® macros called Microarray Calculation and Visualization (MarC-V). MarC-V serves as a flexible tool to be used in the analysis of single-microarray experiments.

### MATERIALS AND METHODS

#### Obtaining and Starting the Program

MarC-V runs on Microsoft Windows® 98, NT®, and 2000® platforms and can be downloaded at no cost at [http://pga.swmed.edu/Information/marcV\\_info.htm](http://pga.swmed.edu/Information/marcV_info.htm). Once downloaded, be sure that support for VBA macros has been installed on your machine. Typically, this is available as an add-in from the Excel "Tools" menu. This may require the Excel installer (Microsoft Office® CD-ROM). Next, double-click on the MarC-V Excel file. A dialog box will appear asking if macros should be enabled. Click on "Enable Macros", and the program will then begin. The opened workbook consists of several worksheets, each pertaining to different functions.

#### Importing Raw Data

Upon opening this Excel workbook, a collection of worksheets is displayed. The default starting worksheet called "Normalization" contains several data columns, along with several buttons. To enter GenePix® results (GPR file),

click on the “Import GenePix Result File” button (Figure 1). The current version of this code can import only GenePix results files. Users should be sure to select one of the two ratio forms (i.e., Cy5/Cy3 or Cy3/Cy5) before importing. This allows for convenient analysis of dye-reversal experiments.

Several data fields are imported into the “Normalization” worksheet from the GPR file. These include mean background and total signal pixel intensities for both channels, gene/open reading frame (ORF) names and accession numbers, spot addresses on the slide, and quality flags. In addition, the entire GPR file is copied into another worksheet of the MarC-V workbook. A more detailed description of GPR file fields is available at [http://www.axon.com/GN\\_GenePix\\_File\\_Formats.html#ResultsFormat](http://www.axon.com/GN_GenePix_File_Formats.html#ResultsFormat) (Axon Instruments, Union City, CA, USA). Once an importable GenePix file is selected from a dialog box, click on the “Open” button to initiate the importing process, data normalization, and calculation of aggregate statistics. The status of this process is displayed in the bottom left corner of the Excel workbook.

Although MarC-V is mainly designed to import GPR files, VBA source code is included and well documented if modification is necessary to import oth-

er file formats. Alternatively, a “custom” MarC-V workbook is available for download (along with a “README” instructions file) from the Web site mentioned above to accommodate data from other microarray image analysis software packages such as ArrayExplorer<sup>®</sup> (6) or ImaGene<sup>™</sup> (BioDiscovery, Marina Del Rey, CA, USA).

### Automated Calculation Flow

For ratio calculation and generation of statistics for microarray datasets, we incorporate the ratiometric method described by Epstein et al. (4). Since the goal is to form normalized expression ratios using one channel divided by the other (Cy5/Cy3 or Cy3/Cy5), some preliminary computation must occur for each element on a microarray. The entire process takes between 30 s and 5 min depending on the CPU speed of the computer and the number of gene/ORF elements that are present on the microarray. For in-house testing, on a PC with an 800-MHz CPU, 128 MB RAM, and 10000 elements, the processing time was 1.2 min. It is recommended that the maximum number of elements does not exceed 25 000, as processing capacity peaks at this point.

First, for each channel, the mean background pixel intensity is subtracted

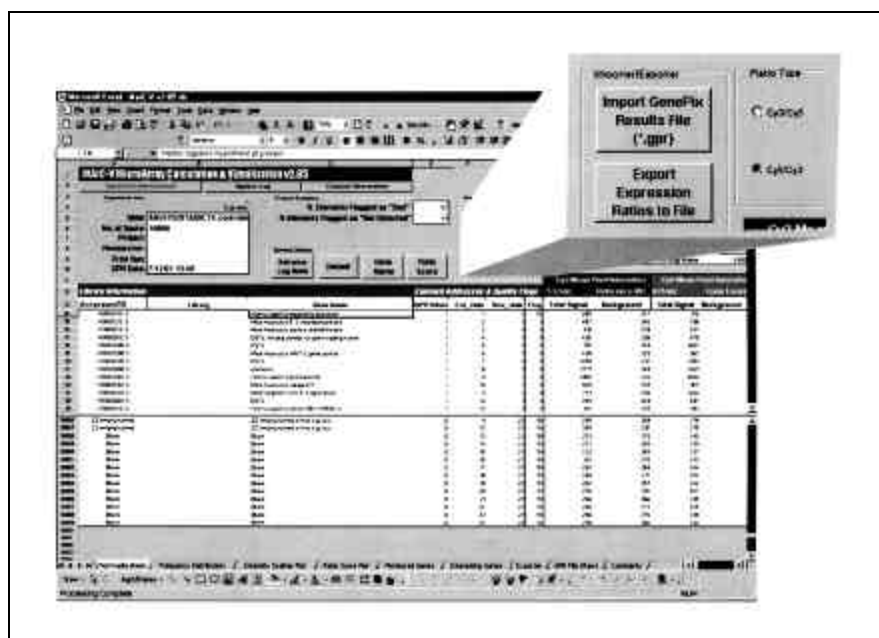
from the total acquired intensity from the GPR file. One problem that results from this calculation is that background intensity values can be less than zero, which eventually contributes to a ratio that is negative or infinite. This situation occurs when faced with very low acquired intensity values, which tend to have substantial uncertainty. The next calculation addresses this issue by calculating a lower bound or threshold such that no background subtracted intensity value can be less than the threshold. The values that are below this threshold are substituted with the threshold value. The threshold is calculated for each channel based on many internally replicated negative controls or blanks, which comprise 1%–2% of the elements. These blanks are typically the resuspension solvents DMSO or SSC. The formula below represents the threshold calculation for each channel, where  $T$ , or threshold, equals the mean of the background subtracted blank intensities  $MB$ , plus a stringency factor  $n$ , multiplied by the average blank SD of both channels  $\sigma$ . The user may change  $n$  to a desired stringency in the “Normalization” worksheet.

$$T_{Cy3} = MB_{Cy3} + n \left( \frac{\sigma_{Cy5} + \sigma_{Cy3}}{2} \right)$$

$$T_{Cy5} = MB_{Cy5} + n \left( \frac{\sigma_{Cy5} + \sigma_{Cy3}}{2} \right)$$

The two resulting thresholds are then adjusted to accommodate for erroneously high blank values caused by spots with foreground pixel intensities lower than background intensities surrounding the spot (dead spots) or very intense artifacts that were not removed using microarray image analysis software before importing. Once the threshold has been applied to all background subtracted measurements, ratios are computed for each element. These ratios are log transformed (base 10) so that equivalent fold changes in expression in the negative or positive direction have the same absolute value.

The last step in this process is ratio normalization. The assumption here is that, for a given array of genes, there will be a subset of expression ratios that will not vary far from 1 (or zero in log space), resulting in a frequency distrib-



**Figure 1.** “Normalization” worksheet after importing data from a GPR file. Zoomed section shows import and export buttons as well as the ratio type option control.

ution that centers most log values around zero. However, the raw data typically show a skewed frequency distribution centered over a non-zero value. Reasons for this skew include differences between the two channels in RNA amount and quality, scanner sensitivity, and efficiency of fluorescent tag incorporation during probe labeling. It is then appropriate to normalize each log ratio by multiplying each denominator by a normalization coefficient such that the sum of all normalized log ratios should be zero. The normalization coefficient is defined as 10 raised to the mean log ratio. Ratios are then scored by intensity and proximity to threshold and pixel saturation. Essentially, this is based on the assumption that ratios formed from higher intensity measurements will have lower uncertainty associated with them (4).

## Viewing Normalized Data

Once normalized data are computed and displayed in the "Normalization" worksheet, users can peruse data columns to gain insight into the results of a particular microarray experiment. Each column has a header and a red triangle in the top right corner of the header cell. Detailed descriptions of a particular data column can be viewed by moving a mouse cursor over this triangle.

It is conventional to sort microarray data to highlight particular properties such as which genes are most differentially expressed. For this reason, buttons are provided for sorting based on particular criteria. Data may be sorted by extreme log ratio (most differentially expressed), gene name, ratio score, or the default based on the original raw data in the GPR file.

Aggregate statistics are also displayed at the top of the "Normalization" worksheet. These include the percentage of elements that were flagged as "bad" or "not found" during image analysis using GenePix software, as well as normalization statistics such as normalization coefficient and normalized mean log ratio. GenePix assigns quality flags of "not found" (-50) automatically if the average pixel intensity for a spot is less than a specified intensity threshold. A "bad" (-100) quality flag is assigned by the user when visually in-

specting each spot for artifacts and other contaminants that would contribute to an inaccurate total signal or background signal. MarC-V eliminates "bad" spots from any statistical calculations.

## Generating Scatter Plots

Another useful feature included with MarC-V is automated plotting of the intensity measurements in both Cy3 and Cy5 channels (Figure 2). To use this feature, click on the "Intensity Scatter Plot" worksheet. Next, generate the scatter plot by clicking on the "Generate Plot" button. This action will copy all normalized data, background subtracted intensities, and the corresponding gene names from the "Normalization" worksheet into a table. Next, a scatter plot is generated with Cy3 measurements on the x-axis and Cy5 measurements on the y-axis. When the scatter plot is complete, ratio outliers may be observed which correspond typically to differential expression ratios that are not in unity with the bulk of the dataset. In addition, cube-root and log intensity scatter plots can be generated

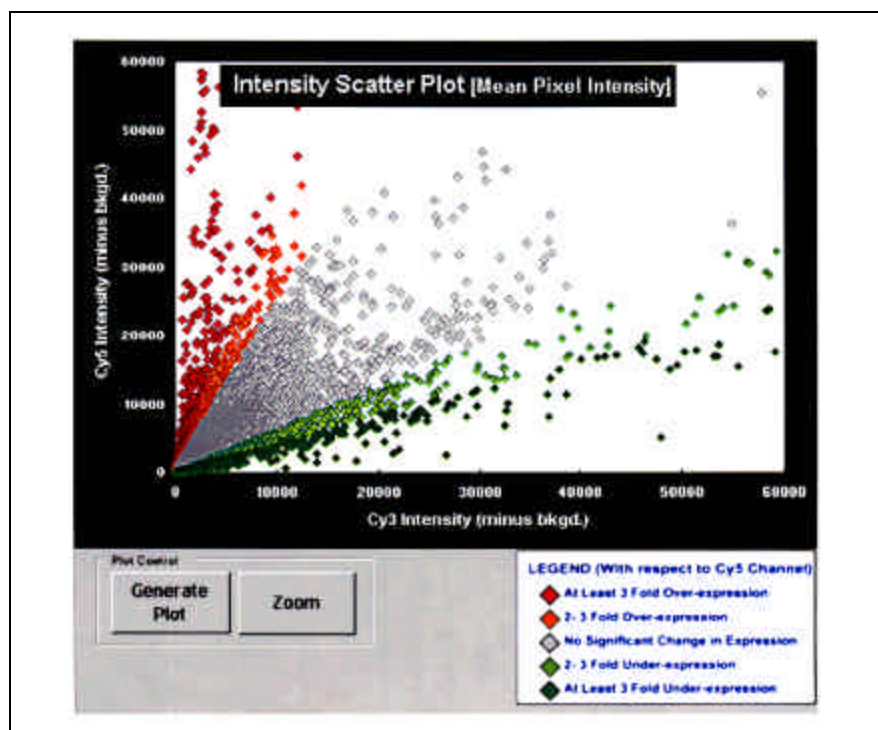
to view gene outliers that are expressed at low levels (9). Each marker on a scatter plot is color-coded based on fold-change in expression after normalization. The values and ranges for this code are defined in a legend just below each scatter plot.

Another scatter plot useful for viewing expression ratio data in MarC-V is the ratio score plot. This scatter plot is color-coded in the same manner as the intensity scatter plots but instead plots log ratios versus the log of the average intensity of both the Cy3 and Cy5 intensity measurements. This plot is generated the same way as the intensity scatter plot.

## Viewing Specific Data Subsets

Because of the volume of data associated with microarray experiments, inspecting subsets of expression data can be very time consuming if done manually. MarC-V has two worksheets included that address this issue.

First, the "Interesting Genes" worksheet displays the top most differentially expressed genes and corresponding



**Figure 2. Scatter plot of microarray data.** Scatter plot depicts Cy3 versus Cy5 background subtracted measurements. Each marker is color-coded to reflect fold change in expression after normalization. Gray markers correspond to elements, which either are flagged unusable or show no significant change in expression.

# MICROARRAY *Technologies*

ratios from the dataset in the “Normalization” worksheet. The number of genes to be displayed may be changed in increments of 50 using scroll buttons at the top of the worksheet. This parameter should be set before populating this table with data. Once this parameter is set, click on the “Extreme log ratio” sorting button in the “Normalization” worksheet. Next, click back to the “Interesting Genes” worksheet to view the results.

A similar worksheet called “Monitored Genes” also creates data subsets based on particular criteria. In this case, individual gene names can be specified, and then each of these names and their associated data can be copied into this worksheet from the “Normalization” worksheet. The resulting dataset will include data from gene names that are present more than once on a microarray. These internal replicates serve as a valuable resource for gauging repro-

ducibility at the gene/clone level.

To use the “Monitored Genes” tool, click on the corresponding worksheet tab. Next, in the “My Favorite Genes” column, type or cut and paste in the gene names that are to be examined. The gene names entered must match the names in the “Normalization” worksheet exactly. Finally, to populate the main table, click on the “Show Me My Genes!” button. This process may take some time depending on how many genes are to be copied and the total number of genes in the “Normalization” worksheet.

## **Exporting Ratios**

Much of the power of microarray expression analysis comes from the ability to generate expression profiles and identify trends in multiple microarray experiments. Several different techniques have recently been applied to analyze multiple experiments, such as hierarchi-

cal clustering (3) and self-organizing maps (8). Software packages such as Cluster (M. Eisen, Stanford University) and J-Express (B. Dysvik, University of Bergen, Norway) (2) that apply these techniques typically take as input a tab-delimited text file consisting of a gene name column and multiple columns of expression ratios. MarC-V has the ability to generate this input file automatically with its exporter function. This exporter can be used by either clicking on the exporter button in the “Normalization” worksheet or by clicking on the “Exporter” worksheet itself. Next, select an experiment number and the array experiment name that will serve as a data column header name describing the experiment in the exported ratios. The experiment number corresponds to which data column the data will occupy in the destination text file. In addition, ratios that are exported can either be normalized or non-normalized, and the gene

1	GENE	15mg 1DAY	15mg 3DAYS	15mg 7DAYS	15mg 10DAYS	3mg 1DAY	3mg 30DAY(B)
2	H34A12	-0.1618	-1.312	0.1677	0.02544	-0.1562	0.8547
3	H33G12	-0.1743	-1.08	0.07954	0.2302	-0.1026	0.8227
4	H35G12	-0.1407	-1.094	0.1668	0.1555	-0.09251	0.8397
5	H37E12	-0.2406	-1.253	0.1154	0.07385	-0.2008	-0.8427
6	H37G6	-0.1719	-1.275	0.1593	0.04162	-0.1548	0.8364
7	H43A12	-0.2461	-1.194	NULL	0.2285	-0.1166	-0.8909
8	H45G6	-0.1361	-1.399	NULL	0.1969	-0.1612	-0.834
9	mitochondrion, cox	-0.04666	-1.368	0.07908	0.07907	-0.1079	0.828
10	mitochondrion, cox	-0.03988	-1.111	0.07585	0.2569	-0.2994	0.8807
11	SM26G6	-0.03208	-0.9004	0.118	0.04162	-0.1218	-0.8147
12	mitochondrion, con	-0.06932	-1.056	0.1084	0.05634	-0.1089	0.8455
13	MITOCHONDRIAL	-0.1326	-1.128	0.1491	0.2337	-0.06107	0.9296
14	MITOCHONDRIAL	-0.201	-1.378	0.1651	0.008911	-0.09718	-0.878
15	21 kd polypeptide	-0.04151	-1.145	0.1151	0.1557	-0.1059	0.8871
16	H33F6	-0.2378	-1.326	0.2132	0.007427	-0.1655	-0.8416
17	H32F12	-0.2726	-1.281	0.1915	0.01868	-0.1423	-0.8449
18	H33H6	-0.1897	-1.265	0.1927	0.08059	-0.1176	0.9169
19	H39B12	0.002224	-1.258	0.1593	0.08383	-0.077	-0.8668
20	H39D6	-0.1239	-1.306	0.2133	0.002829	-0.126	-0.8928
21	H36F6	-0.2488	-1.343	0.1688	0.1729	0.09444	-0.8617
22	H37F12	-0.1882	-1.329	0.1523	0.09717	-0.09459	0.9122
23	H41D6	-0.2006	-1.329	0.1796	0.02382	-0.06298	-0.9084
24	H40F12	-0.2438	0.9602	0.1101	0.04162	-0.115	-0.8777
25	H40H6	-0.2983	-1.301	0.1564	0.05238	-0.1752	0.8804
26	H47F12	-0.07369	-1.312	0.1265	0.04162	-0.03664	0.8611
27	H47H12	-0.03795	-1.409	0.1199	0.04162	-0.08264	-0.8546
28	H50H12	-0.08344	-1.316	0.09945	0.1017	-0.1092	-0.8367
29	H55F12	-0.05757	-1.019	0.1245	0.04162	-0.2036	0.8555
30	H55G6	0.1981	0.9788	0.09081	0.04162	0.09081	0.1487

Figure 3. Sample tab-delimited text file generated by the MarC-V exporter. This file contains six experiments formatted for cluster analysis. Normalized log ratios have been exported.

name column can either be gene names or accession numbers.

For example, if a user selects experiment number 1 with the array experiment name as "15 mg 1 DAY", then the gene name column from the "Normalization" worksheet will be exported to the first column of a new text file with the corresponding ratio data in the second column with a column header called "15 mg 1 DAY". The user will be asked what to name the new file and where to save the file in tab-delimited text format. If experiment number 2 is selected, then the exporter would ask which file to append this ratio data, then export the ratio data to the third column of the previously created text file. Since MarC-V is used to analyze single experiments, several MarC-V workbooks should be opened and data exported sequentially to generate a full, multi-experiment input file for use in clustering software. Figure 3 shows an example of a tab-delimited text file generated by the exporter.

## DISCUSSION

MarC-V is available for Microsoft Windows and Macintosh® operating systems. However, many features are

not available in the Macintosh version because of the processing constraints of Excel. Users are free to modify or add other macros for extending the functionality of the code as they see fit. Thus, other methods such as intra-array variability analysis components similar to those described by Cheng and Wong (1) and Tusher et al. (9) may be incorporated.

In summary, we have described a spreadsheet-based tool called MarC-V for the analysis, normalization, and visualization of cDNA microarray data. There are several worksheets included that all perform separate functions to enable expeditious viewing of particular results that are not easily obtained from sifting through vast microarray datasets manually. In addition, raw data may be imported from GPR files, and processed expression ratios may be exported for use in clustering programs.

## ACKNOWLEDGMENTS

This work was supported with funding for a Program for Genomic Applications (grant no. 5U01HL6688002 to H.R.G.) from the National Heart Lung and Blood Institute and The Donald W. Reynolds Foundation. The authors

wish to thank D. Mittleman, W. Hale, and C. Epstein for valuable comments and contributions, as well as the UT Southwestern Microarray Core Facility.

## REFERENCES

- Cheng, L. and W.H. Wong. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98:31-36.
- Dysvik, B. and I. Jonassen. 2001. J-Express: exploring gene expression data using Java. *Bioinformatics* 17:369-370.
- Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863-14868.
- Epstein, C., W. Hale, IV, and R.A. Butow. 2001. Numerical methods for handling uncertainty in microarray data: an example analyzing perturbed mitochondrial function in yeast. *Methods Cell Biol.* 65:439-452.
- Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J.E. Hughes, E. Snerrud et al. 2000. A concise guide to cDNA microarray analysis. *BioTechniques* 29:548-562.
- Patriotis, P.C., T.D. Querec, B.N. Gruver, T.R. Brown, and C. Patriotis. 2001. Array-Explorer®, a program in Visual Basic for robust and accurate filter cDNA array analysis. *BioTechniques* 31:862-872.
- Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with complementary DNA microarrays. *Science* 270:467-470.
- Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907-2912.
- Tusher, V.G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98:5116-5121.

Received 28 September 2001; accepted 10 December 2001.

### Address correspondence to:

Dr. Jeffrey J. Schageman  
The University of Texas  
Southwestern Medical Center  
Mcdermott Center  
for Human Growth and Development  
and Ryburn Cardiology Center  
5323 Harry Hines Blvd.  
Dallas, TX 75390-8591, USA  
e-mail: Jeff.Schageman@UTSouthwestern.edu

For reprints of this or  
any other article, contact  
Reprints@BioTechniques.com