Understanding Microarray Data

A Guide to Help Users Explore Data and Get Results

Vanderbilt Microarray Shared Resource
Vanderbilt University
Medical Research Building III, Room 9274
465 21st Avenue South
Nashville, TN 37232
(615) 936-3003
(615) 936-3002 (fax)

Table of Contents

General information

Before a lab can submit samples to the VMSR, our Director, Dr. Shawn Levy, must have a brief chat with the lab's Principal Investigator. This can be done by phone or email as well as in person.

Files

For each array experiment that is done, five files are generated. Each file will be named identically, except for the extension of the file. An example of our naming convention is below.

2003-06-14-NIA20K-P8-G3-slide-75.tif

The date that the array was scanned is the first part of the file name. Next, the clone set is given, for example: M5K (mouse 5K), Hu11K (Human 11K), or NIA20K. Following that is the print run: P2 (print 2), P3 (print 3), etc. The term after the print number is the generation number: G2 (generation 2), G3 (generation 3). The last part of the file name indicates the slide number that was used (slide 75). The extension is last and gives the type of file. For each array done, the following files will be generated:

Extension	Description		Comments
.TIF	Image file	Picture of the scanned array	Can be viewed in Photoshop, but since it is a 16-bit image, it will appear in black and white
.GPR	Tab-delimited text file	The results file (ratios, intensities, gene names)	Can be opened in Excel to manipulate data
.JPG	Image file	Shows the array image with both channels overlaid	Used only by the software
.JP1	Image file	Shows one channel scanned (red or green)	Used only by the software
.JP2	Image file	Shows the remaining channel scanned	Used only by the software
.GPS	Settings file	Shows the grid overlaying the image	Can be viewed by opening the settings

Additional information about these file types can be found at: http://www.moleculardevices.com/pages/software/gn_genepix_file_formats.html. In addition to the five files generated from each array experiment, there is another file, in PDF format, for each experiment. This file contains information on the RNA received, the efficiency of dye incorporation, amount of labeled cDNA, etc. In addition, the middle table in this file shows the samples used for each array, and the corresponding slide number.

Accessing Microarray Data

To get his/her data, a user will need to login to the vmsr website (www.vmsr.net) and download the data from their project page. Data is listed under "Results/Raw Data", but additional files may

be found under “Results/Experimental Records”, such as the RNA Workbook, which provides information on sample and hybridization performance. It also provides a table that matches samples with slide numbers.

Installation of the Software Used in Analysis

All arrays are scanned and analyzed (on a basic level) by a program named GenePix, which is available by downloading a trial version from the Molecular Devices website (http://www.moleculardevices.com/product_literature/family_links.php?familyid=14). This version is a demo version and does not have full functionality like the version used in the VMSR. However, the demo version is more than adequate for initial data analysis. GenePix **cannot** be installed on a Macintosh, only on an IBM-compatible personal computer. To install it, download it to the desktop, and double-click the icon. It will then install itself. In most cases, Administrator privileges are required to install software on a PC. Check with the local system administrator if it does not install smoothly. GenePix does have some limitations with regards to high-level data analysis—for example, it cannot be used to view results from multiple experiments simultaneously, nor can it normalize data. For information on other software packages available in the VMSR that will perform high-level data analysis, please see the section titled “VMSR Resources” below.

How to Get Help

This manual aims to provide users with all the knowledge needed to successfully view and understand their microarray data. If, after reading the manual thoroughly and using GenePix’s Help feature, users still need assistance, VMSR staff members are available on a limited basis to assist them. See the “Contact the VMSR” section below for phone numbers and email addresses of the VMSR staff.

VMSR Resources

Scanner Access

The scanner in the VMSR is available for users wishing to scan their own slides. This is done by appointment only. Users may sign up on the VMSR website.

Workstation Access

If a user does not have access to a PC, he or she may use one of the workstations in the VMSR to view his or her data. Again, this is done by appointment only. There is a calendar located on the VMSR website that may be used to secure time on a workstation. Be sure to bring the data on a CD or Zip disk with enough room to save additional files to it (such as analysis results).

Additional Analysis Software

The VMSR currently has other packages to further analyze data: GeneSpring and GeneTraffic are two of them. These can be used to view results from multiple experiments simultaneously, track certain genes across experiments, and perform cluster analysis. After a user has done basic analysis using GenePix or other software (such as Excel), these options can further explored.

Contact the VMSR

To schedule an appointment for help with data analysis, please call the VMSR at 615-936-3001.

Making Sense of Microarray Data

This section will explain, in detail, how to view microarray data and use the analysis tools available within the GenePix software. The best understanding, however, comes with hands-on experimenting with the software. GenePix offers a thorough Help feature, accessed through the last button on the right panel of the screen.

Opening Your Files

1. Open the GenePix Pro software by double-clicking it. The following screen should appear:

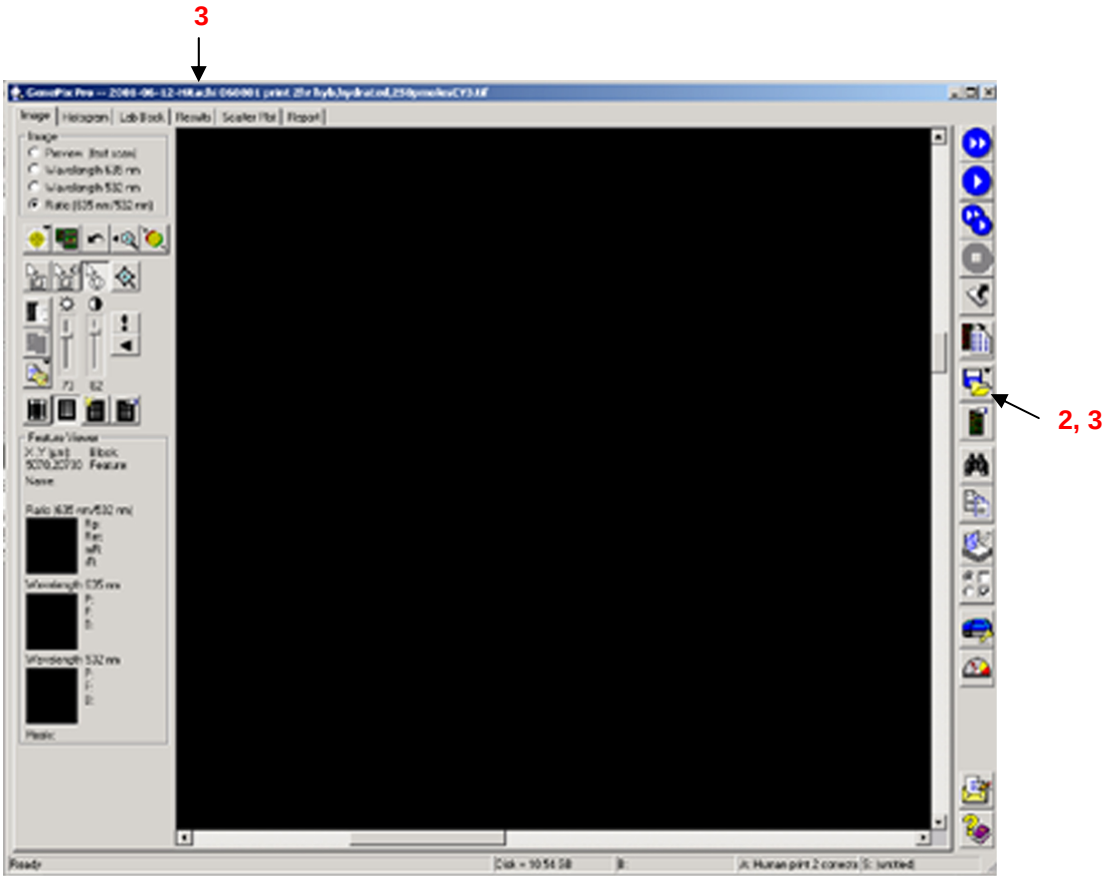

Figure 1. Screen shot of the Image Tab in GenePix. Numbers pointing to the figure indicate the buttons used in each numbered step of these instructions.

2. First, open the **.TIF** file by clicking on the **"***Save*" button on the right of the screen, and selecting *"Open Images"*. A standard "Open File" window will appear. Using the drop-down arrow along the top of the window ("Look in"), select the drive where the data was saved. (If it was saved on the hard drive, the drive is likely **C:** or **E:**. If it was saved on a

Zip disk or CD, the drive is likely **D:** or **F:**.) Once the contents of the correct drive appears in the large window, find the data by double-clicking on the appropriate folder(s). When the data is displayed in the large window, highlight the desired **.TIF** file, and click **"*Open*"**. An image of the array should load.

3. In order to see all the values associated with each spot of the array, the results file (**.GPR**) file must be opened. To open the **.GPR** file, click again on the "*Save*" icon to the right of the screen. Select "*Open Results*". To move between the Image files and the .GPR files, use the navigational tabs at the top of the screen. The "*Results*" tab holds the .GPR file, while the "*Images*" tab contains the .TIF file. **Please note that GenePix software is backwards-compatible, i.e data scanned in GenePix 4.1 can be opened in a more recent version of the software, but data scanned in GenePix Pro 6.0 cannot be opened in a previous version of the software.**
4. Alternatively, one may open their .GPR file and any corresponding image files in one step. Click on the "*Save*" icon and choose "*Open Results*". Make sure that the box labeled "*Open associated images*" is checked. Find the experiment of interest, and click on "*Open*". This will open the **.TIF**, **.JPG**, **.JP1**, and **.JP2** files when the **.GPR** file is opened. These files do not need to be opened by the user.
5. To view data from a different experiment, repeat the steps above.

Exploring the Data

Important Information to Know Before Starting

Each spot on the array is called a feature by GenePix. There are two dyes used in VMSR microarray experiments—Cy5 and Cy3. Cy5 is scanned at wavelength 635nm and fluoresces red. Cy3 is scanned at wavelength 532nm and fluoresces green. Throughout the GenePix application, intensities of spots are not referred to by their color or dye name, but instead by the wavelength at which they are scanned.

Each feature has five ratio values associated with it, *"Ratio of Medians", "Median of Ratios", "Ratio of Means", "Mean of Ratios",* and *"Rgn Ratio*". These are the values that show the level of overexpression or downregulation for each feature. (To learn the differences between these values, see the Appendix or use the GenePix Help feature.) Unless a user requests otherwise, the VMSR calculates all ratios as

$$\frac{\text{Intensity of Cy5 signal}}{\text{Intensity of Cy3 signal}} = \frac{\text{Red}}{\text{Green}}$$

For example, for a particular feature, if the intensity of the Cy5 signal was 40000, and the Cy3 signal intensity was 20000, the ratio would be calculated as 40000/20000 = 2. In other words, the expression of the feature was twice as high in the Cy5-labeled sample than in the Cy3-labeled

sample. It could be said that the particular gene/EST was down-regulated two-fold in the Cy3-labeled RNA sample, or that the gene was two-fold overexpressed in the Cy5-labeled sample.

The same logic applies when the intensity of the Cy3 signal is greater than the intensity of the Cy5. The ratio would still be calculated the same (Cy5/Cy3), but the result would be a number less than 1 (20000/40000=0.5). In this case, the particular gene/EST was down-regulated two fold in the Cy5-labeled RNA sample (or the gene was two-fold overexpressed in the Cy3-labeled sample). To get a better sense of the expression levels of features that show ratios less than 1, just take the inverse of the number (for example, 1/0.5=2, 1/0.25=4, etc).

Using the Results and Scatter Plot screens

1. Look through the **.GPR** file (*Results* tab). Notice that it is just a large text file. Each row corresponds to one feature on the array. Each column contains different pieces of data that define that feature. Scroll sideways through the columns and look at the column headings. Any heading that contains an "F" refers to a feature, and one that contains a "B" refers to the background. Likewise, a column containing "635" refers to Cy5 (red channel), and "532" refers to Cy3 (green channel). Most of these columns can be ignored for the time being, however, look for the columns called *"Ratio of Medians*", *"Median of Ratios", "Ratio of Means*", *"Mean of Ratios*", and *"Rgn Ratio*". These are the values that show the level of overexpression or downregulation for each feature (see explanation above). See the Appendix for a description of all the columns in the **.GPR** file.
2. After becoming familiar with the contents of the **.GPR** file, switch to the Scatter Plot screen by clicking the tab at the top. The X-axis should plot the values from the "F532 Median – B532" column, and the Y-axis should plot the values from the "F635 Median – B635" column. If those values are not the ones plotted (check for text along each axis in order to tell), change the axes to these values by using the drop-down lists on the left panel. The top drop-down list is for the X-axis. Click the arrow on the left of the list and highlight "F532 Median – B532". Follow the same procedure for the Y-axis (bottom drop-down list). Good data will generate a scatter plot similar to the one in Figure 2. If the data does not seem to use the entire grid, click the "Autoscale" button on the left panel (beneath the Y-axis drop-down list). A word of caution: data points that are in the lower left-hand corner (near 0,0, or in the first "block" of the scatterplot) are not usually valid points. These features, in general, have poor morphology and a very low degree of hybridization. These data points should usually be excluded from analysis when running Reports (explained later). Unfortunately, the demo version of GenePix does not allow users to easily exclude these points, but the fully functioning version (available in the VMSR) does (see *Additional GenePix Functionality Available in the VMSR,* below).

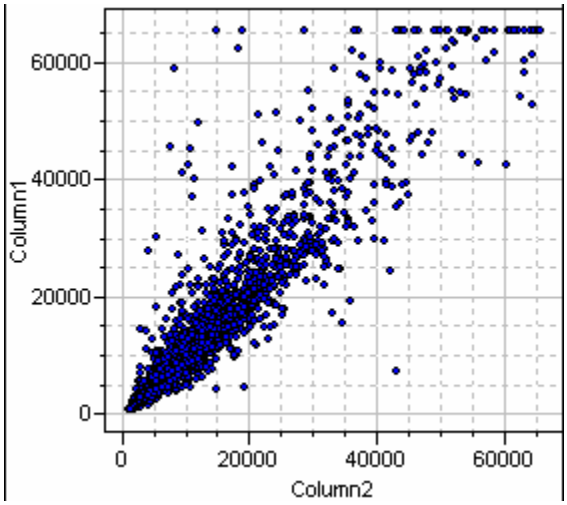

Figure 2. Scatter plot of data.

Each point on the graph corresponds to a feature on the array. Hover over a feature and information about that feature is shown on the bottom left panel, such as the Ratio of Medians value (Rm), and an up-close view of the feature.

Notice that the majority of the points fall along an easily seen "line" extending at a 45° angle from the intersection of the X- and Y-axes. The slope of this line is 1.0, which makes sense. On this graph, the slope is calculated as

$$\frac{\text{Intensity of Cy5 signal}}{\text{Intensity of Cy3 signal}}$$

Notice that this is the same calculation for ratio values. In a microarray experiment, the majority of the features will show no difference in expression between the two samples. In other words, the intensities will be equal, and the ratio for those spots would be 1.0. When graphed, the majority of these features will fall along the line with a slope of 1.0. **Those features that do not fall on this line are differentially expressed between the samples and warrant further investigation.**

Filtering the Data Using the Interesting Features Report

1. Initially, filtering the data is most easily done by using the *Reports* tab. Switch to that screen by clicking the tab at the top. The Reports screen is a web browser. Its scripts are written using VBScript, and all the reports are written using HTML. This makes a particular report easy to save using the Export button (on the left panel)—it will save as an **.HTML** document with all the images saved individually as **.IMG** files.
2. In the scatter plot page (click on tab at top of screen), you will select the data in the very lower left hand corner to flag as "bad" data. These spots carry very low intensity and are mainly either spots that didn't hybridize well or spots that consist of mostly background

noise. Click, drag, and release to highlight the genes that you want to exclude from analysis. After these spots are highlighted yellow, right-click and chose “Flag Bad” from the pop-up menu. These spots will now be excluded from your search for interesting features. Please note that these spots flagged “bad” may still contain useable data. Flagging them bad is the only way GenePix can generate a reasonably-sized list of genes. If the low-intensity spots were left in the analysis (i.e. not flagged bad), the gene list sizes would be in the thousands. For a more advanced analysis, other software packages should be used that can include these lower-intensity genes.

3. Under the Reports tab, click on Report 6 (Interesting Features) under *Analysis Reports*. The Interesting Features Report will filter the data and display only the most differentially expressed spots. This page can be queried by whichever parameters the user desires.
4. To begin, make sure that the first line reads, “Flag features ‘good’ that fulfill the following condition:” and that the second line denotes that all fields should be joined by ‘and’ (as opposed to ‘or’). You may add queries by clicking on the “New Row” button to the right.
5. The suggested way to isolate differentially expressed genes is to set three different fields. The up-regulated genes will be calculated individually of the down-regulated genes. First select “Ratio of Medians (635/532)” to either greater than or less than the ratio cut-off value of your choice (a starting place may be ratios greater than 2.5 and less than than 0.4.). In the second field (click on “New Row” to add fields), set it to “Name $\neq$ 0”. Finally, the third field should read “Flags = 0”. This ensures that only spots with a certain ratio value are shown (demonstrating an expression difference between the Cy5 and Cy3 samples), that the name of the spot is shown (you can also set other fields ‘not equal to zero’ in order to include that information in the final result table), and that all of the spots present are spots that have not been flagged bad.
6. Make sure the box “Construct summary table” is the only one checked. Then click on the “Find” button. This will construct a table of features that meet the selection criteria to be displayed on the page.

To save a report, use the Export button on the upper left panel. A word of warning: the document is saved as an **.HTML** file, and all the images--**each feature shown--**on the report are saved separately as **.IMG** files. Therefore, on a report showing 500 features, one .**HTML** document and 500 **.IMG** files will be saved. To prevent this from becoming difficult to manage (i.e. on one zip disk, 5 **.HTML** pages and 500 **.IMG** files, with no distinction between them), make a new, separate folder for each Interesting Features Report that is saved. For example, create a folder on the zip disk called “Interesting Features”. Within that folder, create new folders called “Slide_60” or “Experiment_1” or however the different Interesting

Features Reports can be designated. Then, as each Interesting Features Report is run, save them in different folders.

Another option is to copy the entire table from the Results tab and paste it into Excel using standard Windows copy/paste functions to allow further manipulation of the data.

Additional GenePix Functionality Available in the VMSR

The GenePix software given to users for their personal use is not fully functional. By using one of the workstations in the VMSR, additional functionality is gained, such as the ability to copy and paste information from GenePix into other software. For example, users can select specific genes on the scatter plot screen by highlighting them, copying the information for those features (using CTRL + C), and then pasting them into Excel. Information from the Reports tab can also be highlighted, copied, and pasted into Excel. In addition, the fully functional version of GenePix allows a user to highlight genes in certain sections of the scatter plot (i.e. the group of genes closest to the 0,0 point on the graph), and exclude them from analysis when running the Interesting Features Report.

Appendix

Column Title	Description
Block	the block number of the feature.
Column	the column number of the feature.
Row	the row number of the feature.
Name	the name of the feature derived from the Array List (up to 40 characters long, contained in quotation marks).
ID	the unique identifier of the feature derived from the Array List (up to 40 characters long, contained in quotation marks).
X	the X-coordinate in µm of the center of the feature-indicator associated with the feature, where (0,0) is the top left of the image.
Y	the Y-coordinate in µm of the center of the feature-indicator associated with the feature, where (0,0) is the top left of the image.
Dia.	the diameter in µm of the feature-indicator.
F635 Median	median feature pixel intensity at wavelength #1 (635 nm).
F635 Mean	mean feature pixel intensity at wavelength #1 (635 nm).
F635 SD	the standard deviation of the feature pixel intensity at wavelength #1 (635 nm).
B635 Median	the median feature background intensity at wavelength #1 (635 nm).
B635 Mean	the mean feature background intensity at wavelength #1 (635 nm).
B635 SD	the standard deviation of the feature background intensity at wavelength #1 (635 nm).
% > B635 + 1 SD	the percentage of feature pixels with intensities more than one standard deviation above the background pixel intensity, at wavelength #1 (635 nm).
% > B635 + 2 SD	the percentage of feature pixels with intensities more than two standard deviations above the background pixel intensity, at wavelength #1 (635 nm).
F635 % Sat.	the percentage of feature pixels at wavelength #1 that are saturated.
F532 Median	median feature pixel intensity at wavelength #2 (532 nm).
F532 Mean	mean feature pixel intensity at wavelength #2 (532 nm).
F532 SD	the standard deviation of the feature intensity at wavelength #2 (532 nm).
B532 Median	the median feature background intensity at wavelength #2 (532 nm).
B532 Mean	the mean feature background intensity at wavelength #2 (532 nm).
B532 SD	the standard deviation of the feature background intensity at wavelength #2 (532 nm).
% > B532 + 1 SD	the percentage of feature pixels with intensities more than one standard deviation above the background pixel intensity, at wavelength #2 (532 nm).
% > B532 + 2 SD	the percentage of feature pixels with intensities more than two standard deviations above the background pixel intensity, at wavelength #2 (532 nm).
F532 % Sat.	the percentage of feature pixels at wavelength #2 that are saturated.
Ratio of Medians	the ratio of the median intensities of each feature for each wavelength, with the median background subtracted.
Ratio of Means	the ratio of the arithmetic mean intensities of each feature for each wavelength, with the median background subtracted.
Median of Ratios	the median of pixel-by-pixel ratios of pixel intensities, with the median background subtracted.
Mean of Ratios	the arithmetic mean of the pixel-by-pixel ratios of pixel intensities, with the median background subtracted.
Ratios SD	the standard deviation of pixel intensity ratios.
Rgn Ratio	the regression ratio.

Rgn R^2	the coefficient of determination for the current regression value.
F Pixels	the total number of feature pixels.
B Pixels	the total number of background pixels.
Sum of Medians	the sum of the median intensities for each wavelength, with the median background subtracted.
Sum of Means	the sum of the arithmetic mean intensities for each wavelength, with the median background subtracted.
Log Ratio	log (base 2) transform of the ratio of the medians.
Flags	the type of flag associated with a feature.
F1 Median – B1	the median feature pixel intensity at wavelength #1 with the median background subtracted.
F2 Median – B2	the median feature pixel intensity at wavelength #2 with the median background subtracted.
F1 Mean – B1	the mean feature pixel intensity at wavelength #1 with the median background subtracted.
F2 Mean – B2	the mean feature pixel intensity at wavelength #2 with the median background subtracted.
Index	the number of the feature as it occurs on the array.