Biological Relevance of GenePix Results
Shawn Handran, Ph.D. and Jack Y. Zhai, Ph.D.
Axon Instruments, Inc. 3280 Whipple Road, Union City, CA 94587


Part I. Introduction

Microarray technology is an effective technology to analyze experimental differences in populations of biomolecules in a high-throughput format. In most cases, microarray scanners use lasers to illuminate one pixel at a time until all the spots on an array chip have been scanned and recorded as a high-resolution image file. The scanned images are analyzed in a data extraction process that measures the relative fluorescence of two fluorophore labels of test and control samples, represented in each spot on the slide. Numerical indicators of the relative binding of test and control substances in each spot are extracted. This involves quantitative processing to calculate various ratios of the fluorescence intensity at two different wavelengths, which often generates a tremendous amount of statistical data that requires further interpretation.

Because of the wide variety in experimental design, this Application note does not make specific recommendations regarding data interpretation. In general, we recommend that if a feature has similar values using these three ratio calculations: Ratio of Medians, Median of Ratios and Regression Ratio, then that feature should be considered trustworthy. This note will provide you with the information you need to make specific decisions about the reliability, quality and biological relevance of your results.

A. Sample Preparation and Generation of Ratio Images

For a quantitative comparison study, the ratio image typically represents the difference in activity of a test sample relative to a reference (i.e., control) sample. The convention used in this application note assumes that the reference sample is labeled with Cy3, and the test sample is labeled with Cy5. The labeled microarray slide is scanned, generating images containing the fluorescence (F) intensity measurements of the Cy3 and Cy5 fluorophores. A ratio image is calculated and displayed as the scan occurs. The GenePix 4000 scanner digitizes and generates a 16-bit unsigned integer image, meaning that the dynamic range for the pixel intensity values for the Cy3 and Cy5 channels range between 0-65535 gray levels. The 16-bit intensity values are used for all the calculations listed in the Results Tab. GenePix Pro allows the user to choose which wavelength (e.g., Cy3 or Cy5) will be the denominator and numerator (Options dialog box, Analysis Tab). For simplicity, this application note assumes that the ratio calculation used is: wavelength 1/wavelength 2 (Cy5/Cy3). In the displayed ratio image (Image Tab), a red spot indicates that the test sample for this feature is expressed at a level higher than the reference. A green spot indicates that the test sample is expressed at a level lower than the reference sample; a yellow spot means that there is similar activity for the two populations. A ratio image therefore provides a visual summary of the whole microarray.

GenePix Pro employs several computational algorithms to calculate ratio values and additional statistical measures, which are listed for each identified printed substance in the Reports Tab. A ratio value of 1.0
indicates that there is no difference between the test and reference. Large changes in ratio values indicate significant differences in activity between the two populations. The results generated by GenePix Pro allow the user to determine whether or not such changes are a reliable measurement of the biological activity of the experiment in question.

B. Histogram

The Histogram Tab in GenePix Pro displays the pixel intensity distribution for the Cy3 and Cy5 channels. The primary function of the histogram is to evaluate PMT settings, to ensure that the voltages are not set too high, which would result in the saturation of pixel intensities to the maximum value of 65535. The Histogram can also be used to balance the signal levels in the Cy3 and Cy5 channels by adjusting the PMT voltages so the red and green histograms overlap. There is more information about histogram settings and image balance in the User’s Guide and on-line Help.

C. Features: Individual Printed Substances Identified by GenePix Pro

GenePix Pro software employs two different types of scatter plots that provide additional visual information about features. The first, known as the Feature Pixel Plot (Figure 1), plots the 635 nm pixel intensities against the 532 nm pixel intensities for a single feature. The Feature Pixel Plot is useful for evaluating the quality of individual features. The Feature Pixel Plot dialog box also displays the median feature and background values and three ratio quantities for ready comparison so that questionable features can be closely inspected and rejected if necessary (e.g., if the three ratio quantities differ by greater than 10%, or if the R2 value is low; see Part II for further discussion).

The second type is displayed in the Scatter Plot Tab (Figure 2), which allows one to plot any numeric result against another. The Scatter Plot Tab is a powerful tool for visually assessing the data, and especially useful in applying “Good” or “Bad” feature flags to data in the Results table. For example, plotting Index vs. Log Ratio provides a rapid visual summary of the changes in the present data set. Part II provides additional examples of the utility of the Scatter Plot.
D. Computing Ratios

GenePix Pro computes five different ratio quantities, which are classified into three groups. The first ratio calculation method uses median or mean values derived from the whole feature (Ratio of Medians, Ratio of Means), whereas the second method calculates ratios pixel-by-pixel (Median of Ratios, Mean of Ratios). A third method extrapolates the ratio value by regression analysis (Regression Ratio). The pixels used to calculate each ratio are shown in Figure 3.

High quality microarray slides will yield similar values for each ratio calculation method, so a comparison of ratio values is a robust way of evaluating whether changes in the relative levels of the control and test populations are significant. The variety of calculation methods provided by GenePix Pro is intended to give the user a complete repertoire of ratios, but no ratio derivation is more “correct” than any other. The following points are considerations you should keep in mind when evaluating ratio results.

1. For microarray analysis, median intensities are generally preferred to mean intensities because the median is less affected by extreme values at either end of the intensity distribution, which typically are caused by dust contaminants or other artifacts.

2. The Regression Ratio differs from the other ratio derivations because it does not use background subtraction. The regression fit can be viewed in the Feature Pixel Plot (Figure 1). Figure 3 shows two different plots of the same feature, the left panel showing the feature pixels in yellow and background pixels in black. The Ratio of Medians and Median of Ratios are derived from these pixels values and are plotted as a red line (Rm) or as a green line (mR). In contrast, the Regression Ratio is calculated from the regression pixels (shown in green; Figure 3, right panel), which are not distinguished as features or background. The Regression Ratio therefore provides an independent ratio derivation for comparison with ratio quantities that employ background subtraction, because the regression ratio includes background pixels when calculating the regression fit, whereas the Ratio of Medians and Median of Ratios subtract a median background intensity value from the respective feature intensities. Note that the regression pixels do not necessarily include all of the feature or background pixels and may also include additional pixels not assigned to either feature or background; this is because the two-pixel exclusion around the feature indicator does not apply in the calculating the regression ratio. The least squares best-fit curve (blue line) is calculated from the regression pixels and the slope is the Regression Ratio.

3. Together, if all three methods (Ratio of Medians, Median of Ratios, Regression Ratio) yield similar values, such a value for a given feature should be considered reliable. Small differences among these three methods do not necessarily compromise the reliability of a feature, but may indicate a spot that requires visual inspection for abnormalities such as contaminates, or may indicate a problem with staining protocol or image acquisition settings.
Part II. Complete Description of GenePix Results and Biological Relevance

Block
Definition: The Block number assigned by GenePix Pro, increasing from left to right and then top to bottom. A Block is the basic organizing unit that consists of a set number of rows and columns of arrayed spots.

Biological relevance: A block corresponds to a single pin in the array printer. A bent, or otherwise defective pin will therefore be manifested in the printing of a single block.

Column/Row
Definition: The Column or Row number of a feature.

Biological relevance: The Column and Row values can be plotted against other quantities in the Scatter Plot to visualize directional trends in array signal; such trends might indicate systematic errors in the microarray slide.

Name/ID
Definition: The user-defined Name and ID number associated with each spot on the microarray. These fields are blank unless a GenePix Array List (GAL) file containing the Name and ID for each spot has been loaded and applied.

Biological relevance: The Name and ID number are usually pre-defined. The name may be assigned to the biomolecule by the original authors, and the ID number is typically an accession number for the biomolecule in a database such as GenBank. By selecting a feature in the Results Tab and pressing the “Go to web” button, the web browser will be invoked and directed to the user-defined genome database. Information regarding the biomolecule in question will be accessed automatically. The ID is the most critical parameter for post-analysis in other software programs because it is the unique identifier associated with the results for that feature. In addition, the ID must be correctly formatted in order to successfully retrieve information from Genomic website databases. If the ID is not formatted to the specifications of the website you have set in the Options dialog box (Analysis Tab), an error will be returned by the website server because the erroneous ID will not be found in the database.

X/Y
Definition: The physical location of the feature on the slide, in microns. The ordinate (0,0) is the upper left corner of the scan area.

Biological relevance: The X and Y coordinates identify the exact location of the feature in question and can be helpful in quickly identifying a feature in the Image Tab. In addition, the X and Y values can be plotted in the Scatter Plot to identify potential systematic error in the microarray slide or hybridization protocol.

Diameter
Definition: The diameter of the feature-indicator ring, in microns.

Biological relevance: The diameter value can be used in the Scatter Plot to evaluate the consistency of the arrayed spots. The variability in feature diameters, if exceeding a set limit (as determined by the user), may be indicative of problems or errors in the microarray printer or glass surface.

F635 Median
F635 Mean
F532 Median
F532 Mean
Definition: The median or mean intensity values at wavelength 1 (635 nm) or 2 (532 nm) of all pixels that fall within the feature-indicator ring.

Biological relevance: These are the median or mean intensity values without background subtraction for the feature in question. As such, they are the “raw data” of the experiment and, through the Scatter Plot, can be plotted against ratio values (or any other quantity) for thorough data analysis. Contaminants (e.g.,
dust or excess dye) would give rise to divergent median and mean values, whereas high quality microarrays would yield similar, nearly identical median and mean values. Thus plotting the median and mean fluorescence intensity values in the Scatter Plot would be a useful quality control comparison.

**F635 SD**

**F532 SD**

**Definition:** The standard deviation of the intensity values at wavelength 1 (635 nm) or 2 (532 nm) of all pixels that fall within the feature-indicator ring.

**Biological relevance:** A measure of the intensity distribution spread for the feature in question. A large SD may be indicative of a technical problem in the microarray itself or in the hybridization protocol (e.g., inadequate or uneven staining). Systematic errors may be detected by plotting the **F635 SD** or **F532 SD** against the block, column or row in the Scatter Plot.

**B635 Median**

**B635 Mean**

**B635 SD**

**B532 Median**

**B532 Mean**

**B532 SD**

**Definition:** The median, mean or standard deviation of the background intensity values at wavelength 1 (635 nm) or 2 (532 nm) of all pixels that meet the criteria to be assigned as background pixels for a given feature.

**Biological relevance:** A useful quality control test is to compare the Mean and Median background values; a discrepancy (e.g., ten percent or greater) would indicate that there is a contaminant or some other aberration in the background of the feature in question. Background intensities can also be a useful indication of microarray quality; e.g., a high background may indicate that inappropriately high PMT voltages were used, or that there is non-specific staining or insufficient destaining.

**% >B635 +1 SD**

**% >B635 +2 SD**

**% >B532 +1 SD**

**% >B532 +2 SD**

**Definition:** The percentage of feature pixels at wavelength 1 (635 nm) or 2 (532 nm) that have intensity values greater than 1 or 2 SD above the median background intensity value.

**Biological relevance:** This parameter measures % of feature pixels, not background pixels, so it does not indicate “dispersion of background pixels” or anything wrong with background. It can be used to threshold features that have a value greater than 1 or 2 SD’s above the background, and therefore have decent signal quality, and are probably reliable values. These values can be used to threshold high or low signal values in features with large changes in activity.

**F635 % Saturated**

**F532 % Saturated**

**Definition:** The percentage of feature pixels at wavelength 1 (635 nm) or 2 (532 nm) that have the maximum 16-bit intensity value of 65535.

**Biological relevance:** Any value greater than zero indicates that the signal from some pixels in the feature exceed the dynamic range of the detection system. True signal intensity cannot be measured accurately from saturated pixels. It may be necessary to re-scan the microarray slide using a lower voltage on the PMT detector for one or both wavelengths to ensure precise ratiometric analysis. Viewing the **Histogram Tab** is the fastest and easiest way to ascertain whether the F635 or F532 intensities are too high.
Ratio of Medians (Rm)

**Ratio of Means**

**Definition:** The 635 nm/532 nm (background-corrected) ratio of the median or mean intensity values, calculated from the whole feature.

**Biological Relevance.** The mean intensity from a region of interest is often used in many different types of quantitative fluorescence imaging applications. Mean intensity values have been traditionally used in quantitative cellular fluorescence imaging; however, for microarrays, which typically have markedly smaller regions of interest, the median is preferable because it is less affected by extreme values, such as particle contaminants, which can be of comparable size to microarray features. The **Ratio of Medians** therefore, would be considered a valid calculation method in extracting intensity measurements from microarray features. An overall summary of the microarray data can be viewed in the Scatter Plot by graphing the **Ratio of Medians** or the **Ratio of Means** against the index number (which is simply, the sequential enumeration of each feature in the microarray).

Median of Ratios (mR)

**Mean of Ratios**

**Definition:** The 635 nm/532 nm ratios (background corrected) are calculated for each pixel within the feature and a median or mean value from these ratios is returned.

**Biological Relevance.** This calculation method provides an alternative method for measuring changes in activity. The ratio values for individual pixels can be viewed in the Image Tab and Feature Viewer (see Figure 4). One advantage of calculating the **Mean of Ratios** is that a **Ratios SD** can be derived, providing an additional way to visualize the spread of the microarray ratio data. High quality microarrays will generate very similar values using both methods of ratio determination (whole features versus pixel-by-pixel calculation). A large discrepancy between the two calculation methods (i.e., Ratio of Medians and Median of Ratios) may indicate quality control concerns with the microarray slide or experimental procedure.

**Figure 4**

Pixel Results in the Feature Viewer
**Legend.** Values indicated by the blue boxes in the Feature Viewer refer to individual pixel values: \( Rp \) is the ratio value for the given pixel (indicated by the cursor); \( P \) is the pixel intensity value for each respective wavelength (635 nm and 532 nm).

**Ratios SD**

**Definition:** The standard deviation (SD) of the intensity ratios of all pixels within a feature. The Ratio SD is derived from ratio values calculated on a pixel-by-pixel basis, with the median background value of the current feature subtracted from each pixel.

**Biological relevance:** The Ratios SD indicates the spread of the pixel-by-pixel Mean of Ratios derivations for a given feature. The Ratios SD plotted against the Index in the Scatter Plot would facilitate the selection and flagging of features based on a user-defined threshold (e.g., Ratios SD greater than 3 could be flagged “Bad”). Additional quantities, such as the Mean of Ratio (or other ratio quantities, Log Ratio, etc.) plotted against the Ratios SD can be used to verify that features with the highest changes in activity also have acceptable (e.g., less than 2 SD) levels of data dispersion.

**Rgn Ratio (rR)**

**Definition:** A third method of calculating the 635 nm/532 nm ratios by plotting the fluorescence intensity values for each pixel against each other (i.e., 635 nm versus 532 nm) and determining the slope of the least squares best-fit regression line.

**Biological Relevance.** The Regression Ratio does not differentiate feature from background pixels, and does not use a background subtraction method; therefore the Regression Ratio provides an independent measure that can be compared to the Ratio of Medians and Median of Ratios calculation methods. The Rgn R is useful in ascertaining individual feature quality because signal spikes caused by saturated pixels or contaminants such as dust affect the Rgn R more than the other ratio calculation methods. If the Rgn R of the entire microarray is markedly different from the \( Rm \) and \( mR \), it is possible that the PMT settings are set too high or there is a lot of dust on the slide.

**Rgn R**

**Definition:** \( Rgn \ R^2 \) is the coefficient of determination for the least-squares regression fit of a given feature, which is obtained from regression pixels used to calculate the Rgn Ratio. It is the square of the correlation coefficient and ranges in value between 0 and 1.

**Biological Relevance:** \( Rgn \ R^2 \) is the proportion of regression pixels that can be described by the regression fit. For example, a \( Rgn \ R^2 \) of 1 indicates that all the pixels are described by the regression fit (i.e., a straight line); a \( Rgn \ R^2 \) value of 0 indicates that none of the data are described by the regression. The \( Rgn \ R^2 \) value can be useful in determining the uniformity of a given feature; features with spotty appearance (which may indicate uneven distribution of deposited sample, or inefficient hybridization) or features disproportionately lacking one or the other fluorophore will have lower \( Rgn \ R^2 \) values. Lower \( Rgn \ R^2 \) values for a large number of features within a block or slide could indicate poor quality spot printing or problems in the hybridization protocol.

**F Pixels (F)**

**Definition:** The number of pixels within a feature. Note that F Pixels refers to the number of pixels comprising the feature, and differs from the F value indicating feature intensity in the Feature Viewer.

**Biological relevance:** F Pixels is the number of “observations” within a feature spot, i.e., the n (number) value in statistical analysis.

**B Pixels**

**Definition:** The number of pixels that comprise the background for a given feature, as defined in the User’s manual. Indicated as ‘B=number’ at the bottom of the Feature Viewer following the value given for F Pixels.

**Biological relevance:** The number of B pixels can be compared among all features in a microarray as a quality control value. Again, the Scatter Plot can be employed for rapid evaluation (e.g., Index vs. B Pixels). The relationship between the number of F pixels and B pixels should remain constant.
(approximately 3-fold), assuming a uniform feature size throughout the microarray; arrays with large B pixel values (more than 3-fold above the expected F Pixel size) indicates feature-indicators that are substantially larger than the nominal diameter. In such cases, the features in question probably have a background abnormality that gave rise to an unusually large feature-indicator and should be interpreted with caution.

**Sum of Medians**

**Sum of Means**

**Definition:** The sum of the background-subtracted median or mean pixel intensity values for both wavelengths.

**Biological Relevance:** A feature with poor signal in both wavelengths will give rise to a relatively small **Sum of Medians** value. When pixel intensity values are near background levels, even small differences in median pixel intensity values between the two wavelengths can give rise to markedly high or low ratio values. Ratio values for features with unusually low **Sum of Medians** values should therefore be interpreted with caution. Like Sum of Medians, the **Sum of Means** can provide another level of interrogation of a feature, although, as mentioned previously, analysis of mean values tends to be less robust than analysis of median values.

**Log Ratio**

**Definition:** The log₂ of the **Ratio of Medians**.

**Biological Relevance:** The **Log Ratio** values gives an index of the fold change in activity. Negative values indicate a decrease in activity between the control and test populations whereas positive values indicate an increase. Each successive integer step indicates a log₂ increase or decrease in the relative activity levels (e.g., log₂(1) = 2, represents a 2-fold change; log₂(2) is a 4-fold change; log₂(3) = 8, etc.).

**F635 Median – B635**

**F532 Median – B532**

**F635 Mean – B635**

**F532 Mean – B532**

**Definition:** The median or mean pixel intensity of wavelength 1 (635 nm) or 2 (532 nm) for the feature in question with the median background of the respective wavelength subtracted.

**Biological Relevance:** These values are especially useful in scripting because the background subtraction operation has already been performed (thereby eliminating the need for additional code that performs the background subtraction in the script). Since the background level has been subtracted, features with unusually high background levels or features that are weakly fluorescent (with pixel intensity values near background) can be differentially affected by the background subtraction; therefore, these values should not be consulted for quality control purposes. It would be more appropriate in such cases to view the **Sum of Medians** or the **F635/532 Median** or **Mean** values.

**Flags**

**Definition:** see Table

**Biological Relevance:** none

<table>
<thead>
<tr>
<th>Feature</th>
<th>User-selectable?</th>
<th>Indicates:</th>
<th>Value in “Flags” column:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>Yes</td>
<td>A feature of interest to user</td>
<td>+100</td>
</tr>
<tr>
<td>Bad</td>
<td>Yes</td>
<td>Feature fails quality criteria set by user</td>
<td>-100</td>
</tr>
<tr>
<td>Absent</td>
<td>Automatic</td>
<td>Entry missing in GAL file</td>
<td>-75</td>
</tr>
<tr>
<td>Not Found</td>
<td>Automatic</td>
<td>Auto-align fails to detect a feature</td>
<td>-50</td>
</tr>
</tbody>
</table>
Part III. Data Normalization.

Normalization is the process of standardizing several sets of data to a uniform scale. Applying a normalization factor can allow several different microarray slides to be quantitatively compared. GenePix Pro calculates several different normalization factors, which are listed in the header information of the GenePix Results (gpr) file, or can be viewed in the Results Tab (this is for display purposes only, the results are not altered by selecting a normalization factor to view). The calculation assumes that the mean ratio value of all features within the microarray is one. The normalization factor returned by GenePix Pro is the value that will scale the data such that the mean ratio value is one. The GenePix Pro User’s Guide has further discussion of normalization factors (Chapter 3).

Alternatively, one might want to normalize a number of features to one or more control genes located on the same microarray slide. In order to normalize GenePix results to one or more control genes on the same microarray slide, the user must calculate the normalization factor from the control features of choice and apply it to the data, either by creating a custom script, or by exporting the data to a spreadsheet program.

Additional Resources:

The “axongenomics” discussion group at: http://groups.yahoo.com/group/axongenomics.

Axon Instruments technical support resources:
Web: http://www.axon.com/mr_Technical_Support.cfm
Phone: 510-675-6200