



## GP3: GenePix post-processing program for automated analysis of raw microarray data

M. R. Fielden, R. G. Halgren, E. Dere and T. R. Zacharewski

Department of Biochemistry & Molecular Biology, National Food Safety & Toxicology Center, Institute for Environmental Toxicology, Michigan State University, East Lansing, MI 48824, USA

Received on October 7, 2001; revised on November 22, 2001; December 19, 2001; accepted on December 21, 2001

### ABSTRACT

**Summary:** Here we describe an automated and customizable program to correct, filter and normalize raw microarray data captured using GenePix, a commonly used microarray image analysis application. Files can be processed individually or in batch mode for increased throughput. User defined inputs specify the stringency of data filtering and the method and conditions of normalization. The output includes gene summaries for replicate spots and descriptive statistics for each experiment. The source code (Perl) can also be adapted to handle raw data output from other image analysis applications.

**Availability:** <http://bch.msu.edu/~zacharet/microarray/GP3.html>

**Contact:** [tzachare@msu.edu](mailto:tzachare@msu.edu)

**Supplementary information:** <http://bch.msu.edu/~zacharet/microarray/GP3.html>

### INTRODUCTION

Image analysis of two-color fluorescent cDNA microarray produces a large number of raw data points describing the spot fluor (e.g. Cy5 or Cy3) intensity, background fluor intensity, and a variety of other spot quality measurements. Typically, the raw spot intensity values for one fluor are corrected for background signal and then compared to the corrected spot intensity value of the other fluor to generate a ratio. This ratio represents the relative difference in gene expression between the two samples co-hybridized on the microarray (e.g. control and test cDNA). However, systematic and experimental biases can exist between the two fluor-labeled cDNA populations, resulting in inaccurate quantitation of relative differences in gene expression. The error is often associated with (i) differences in the efficiency of incorporation of fluor-labeled nucleotides into the cDNAs by reverse transcriptases, (ii) differences in the stability and fluorescence emission characteristics of the fluors, and (iii) differences in RNA loading, quality and

sample handling. As a result, signal intensity values for each fluor are often normalized, or transformed, by a correction factor. This mathematical correction attempts to remove systematic and experimental biases in fluor characteristics so that accurate ratios can be calculated (Schuchardt *et al.*, 2000; Yang *et al.*, 2001; Tseng *et al.*, 2001). Furthermore, failure to remove spots below threshold levels (i.e. no expression) or at saturating levels can lead to invalid or undefined ratios.

Data filtering, correction, and normalization of raw data to produce valid ratios are often performed manually using spreadsheets, which can be time consuming and prone to error. Summarizing the results from experiments and tracking replicate spots and errors can also be very time consuming. In order to automate and increase the throughput of processing raw microarray data, while simultaneously minimizing human intervention, a script was developed to automatically process raw microarray data from GenePix image analysis software (Axon Instruments, Union City, CA).

### ALGORITHM

The algorithm uses a threshold to define detectable expression in order to filter those spots that are considered below the limits of detection and not appreciably different from background values. A spot is considered below the limit of detection if

$$S_{ij} < B_{ij} + x\sigma_{B_{ij}} \quad (1)$$

where  $S_{ij}$  is the median spot signal intensity for gene  $i$  ( $i = 1, \dots, n$  genes on the array) in channel  $j$  ( $j = 1$  or  $2$ ),  $B_{ij}$  is the median background spot intensity for gene  $i$  in channel  $j$ ,  $x$  is a user defined threshold (default = 3), and  $\sigma_{B_{ij}}$  is the standard deviation of  $B_{ij}$ . If the spot is flagged in one channel but not the other,  $S$  can be set to a user-defined baseline value to avoid undefined ratios (default raises  $S$  to the threshold level for that gene according to equation (1)). A flag is set in the output file to indicate

\*To whom correspondence should be addressed.

that the calculated ratio may be inaccurate as one channel was below the limits of detection. If the spot is flagged in both channels, the gene is removed from any further analysis and no valid ratio is calculated. Spots are also flagged if  $S$  is saturated (i.e.  $S = 65\,536$  for GenePix) in either channel. This indicates that the calculated ratio may be inaccurate since the true value for  $S$  is unknown. If  $S$  is saturated in both channels, the gene is removed from any further analysis and a valid ratio is not calculated. Spots are corrected for background signal to produce a corrected spot signal intensity ( $S'$ ) according to equation (2).

$$S'_{ij} = S_{ij} - B_{ij} \quad (2)$$

Corrected spot signal intensities ( $S'$ ) are normalized by one of two linear normalization methods as defined by the user input. This step scales the distribution of log ratios closer to a mean of zero, such that the distributions of intensities are equivalent and comparisons between channels are more accurate. Intensity values are normalized in log space in order to make normalization additive and to make the variation in intensity less dependent on absolute magnitude. The two normalization options are termed (1)  $z$ -score normalization, and (2) global normalization.

The  $z$ -score normalization is a linear transformation applied to the  $\log_2 S'$  values so that the distribution of  $z$ -score normalized values has zero mean and unit variance for that channel. This is done by scaling the  $\log_2$  signal intensity ( $\log_2 S'$ ) of each spot on the array by subtracting the mean of all ( $n$ )  $\log_2 S'$ , or a subset ( $n - m$ ) of  $\log_2 S'$ , and dividing by its standard deviation (equations (3) and (4)). The scaled result is inverse transformed and termed the normalized signal intensity ( $N_{ij}$ )

$$N_{ij} = 2^{(\log_2(S'_{ij}) - X_j)/\sigma_{X_j}} \quad (3)$$

where  $X_j$  is

$$X_j = \frac{\sum_{i=m}^n \log_2(S'_{ij})}{n - m} \quad (4)$$

and  $\sigma_{X_j}$  is the standard deviation over the same set of spots used to calculate  $X_j$ .

A subset of  $\log_2 S'$  may be appropriate when values at either end of the distribution may inappropriately bias the scaling factor. This may occur when comparing two very distinct tissue types with divergent expression profiles. Only spots that are not flagged are included in the calculation of  $X$  and  $\sigma_{X_j}$ . To exclude outliers at either end of the distribution of  $\log_2 S'$ , a trimmed  $X$  (i.e. a subset of  $\log_2 S'$ ) and its standard deviation can be calculated as defined by the user input. By default, a 90% trimmed mean of valid spots is calculated for  $X$ , such that 5% of the values at either end of the distribution are excluded from  $X$ . This assumes approximately 90% of the genes on an array will be unchanged by treatment. The value chosen

will depend on the expected degree of variation between the two samples being compared. More divergent samples may require a smaller subset (i.e. 50%).

The global normalization method scales the  $\log_2$  signal intensity ( $\log_2 S'$ ) of each spot on the array by subtracting the mean of all ( $n$ )  $\log_2 S'$ , or a subset ( $n - m$ ) of  $\log_2 S'$  (equation 5).

$$N_{ij} = 2^{\log_2(S'_{ij}) - X_j} \quad (5)$$

Regardless of the method used to normalize, both the normalized ratios ( $R$ ) and  $\log_2$  transformed ratios ( $R'$ ) of channel 1 and channel 2 are calculated for gene  $i$  according to

$$R_i = \frac{N_{i1}}{N_{i2}} \quad (6)$$

$$R'_i = \log_2 R_i \quad (7)$$

The results of data filtering, correction, and normalization are appended to the raw data in a comma separated values (CSV) file to facilitate graphing and visualization of the results and to preserve the original raw numbers. The CSV files can be opened in a spreadsheet program, such as Microsoft Excel. The file includes flag fields for spots that do not pass the threshold criteria in channel 1 or 2, as defined by equation (1), as well as a flag for spots that were saturated in channel 1 or 2. This is used to judge the accuracy of the ratio measurements.  $\log_2 S'$ ,  $N$ ,  $R$ ,  $R'$  are recorded for all valid spots on the microarray. The geometric mean of the signal intensity ( $S$ ) in both channels is also calculated (denoted  $G$ ) and appended to the results, in addition to  $\log_2 G$  and its percentage of maximum ( $100 \times G/65536$ ). This is used primarily for graphing and estimating expression levels.

A second CSV file is produced to summarize the normalized signal intensity values (arithmetic mean, standard deviation, and coefficient of variation) for replicate spots on the array. The average signal intensity data (i.e. average of  $G$ ,  $\log_2 G$  and its percentage of maximum across replicates) are also included in the summary file.

A descriptive file is produced for each microarray to summarize the results of the experiment. This includes the header information from the GenePix results file and the user-defined parameters selected for the analysis. Descriptive statistics and a summary of the experimental results include a summary of the flags, normalization factors, correlation between channels, average spot and background signal intensities, and the distribution and percent distribution of valid ratios.

## IMPLEMENTATION

The Perl-based script is available for download at <http://bch.msu.edu/~zacharet/microarray/GP3.html>. The program was written in Perl 5.6.0 and requires the

external Perl module Statistics::Descriptive v. 2.3 (authored by Colin Kuskie). The module is publicly available at [www.CPAN.org](http://www.CPAN.org). Although the program can currently accommodate only GenePix result files, the available source code can be easily modified to accommodate a variety of file formats. The script can be executed to process single files or entire directories for increased throughput. Documentation for use can be found at <http://bch.msu.edu/~zacharet/microarray/GP3.html>.

Due to the wide use of GenePix as a microarray image analysis application, as well as the necessity for a simple, intuitive and effective means of processing raw microarray data in a fast and customizable manner, we expect researchers using microarrays to find this script

of great utility. It is also expected that researchers using other image analysis applications will modify the existing source code to accommodate other output file formats.

## REFERENCES

- Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. and Herzog, H. (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res.*, **28**, E47.
- Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. and Wong, W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549–2557.
- Yang, W.H., Dudoit, S., Luu, P. and Speed, T.P. (2001) *Normalization for cDNA microarray data*, SPIE BiOS 2001, San Jose, CA.