Microarray data normalization and transformation

John Quackenbush

doi:10.1038/ng1032

Underlying every microarray experiment is an experimental question that one would like to address. Finding a useful and satisfactory answer relies on careful experimental design and the use of a variety of data-mining tools to explore the relationships between genes or reveal patterns of expression. While other sections of this issue deal with these lofty issues, this review focuses on the much more mundane but indispensable tasks of 'normalizing' data from individual hybridizations to make meaningful comparisons of expression levels, and of 'transforming' them to select genes for further analysis and data mining.

The goal of most microarray experiments is to survey patterns of gene expression by assaying the expression levels of thousands to tens of thousands of genes in a single assay. Typically, RNA is first isolated from different tissues, developmental stages, disease states or samples subjected to appropriate treatments. The RNA is then labeled and hybridized to the arrays using an experimental strategy that allows expression to be assayed and compared between appropriate sample pairs. Common strategies include the use of a single label and independent arrays for each sample, or a single array with distinguishable fluorescent dye labels for the individual RNAs. Regardless of the approach chosen, the arrays are scanned after hybridization and independent grayscale images, typically 16-bit TIFF (Tagged Information File Format) images, are generated for each pair of samples to be compared. These images must then be analyzed to identify the arrayed spots and to measure the relative fluorescence intensities for each element. There are many commercial and freely available software packages for image quantitation. Although there are minor differences between them, most give high-quality, reproducible measures of hybridization intensities.

For the purpose of the discussion here, we will ignore the particular microarray platform used, the type of measurement reported (mean, median or integrated intensity, or the average difference for Affymetrix GeneChips™), the background correction performed, or spot-quality assessment and trimming used. As our starting point, we will assume that for each biological sample we assay, we have a high-quality measurement of the intensity of hybridization for each gene element on the array.

The hypothesis underlying microarray analysis is that the measured intensities for each arrayed gene represent its relative expression level. Biologically relevant patterns of expression are typically identified by comparing measured expression levels between different states on a gene-by-gene basis. But before the levels can be compared appropriately, a number of transformations must be carried out on the data to eliminate questionable or low-quality measurements, to adjust the measured intensities to facilitate comparisons, and to select genes that are significantly differentially expressed between classes of samples.

Expression ratios: the primary comparison

Most microarray experiments investigate relationships between related biological samples based on patterns of expression, and the simplest approach looks for genes that are differentially expressed. If we have an array that has N_{array} distinct elements, and compare a query and a reference sample, which for convenience we will call R and G, respectively (for the red and green colors commonly used to represent array data), then the ratio (T) for the ith gene (where i is an index running over all the arrayed genes from 1 to N_{array}) can be written as

$$T_i = \frac{R_i}{G_i}.$$

(Note that this definition does not limit us to any particular array technology: the measures R_i and G_i can be made on either a single array or on two replicate arrays. Furthermore, all the transformations described below can be applied to data from any microarray platform.)

Although ratios provide an intuitive measure of expression changes, they have the disadvantage of treating up- and downregulated genes differently. Genes upregulated by a factor of 2 have an expression ratio of 2, whereas those downregulated by the same factor have an expression ratio of (–0.5). The most widely used alternative transformation of the ratio is the logarithm base 2, which has the advantage of producing a continuous spectrum of values and treating up- and downregulated genes in a similar fashion. Recall that logarithms treat numbers and their reciprocals symmetrically: $\log_2(1) = 0$, $\log_2(2) = 1$, $\log_2(^1/_2) = -1$, $\log_2(4) = 2$, $\log_2(^1/_4) = -2$, and so on. The logarithms of the expression ratios are also treated symmetrically, so that a gene upregulated by a factor of 2 has a $\log_2$(ratio) of 1, a gene downregulated by a factor of 2 has a $\log_2$(ratio) of –1, and a gene expressed at a constant level (with a ratio of 1) has a $\log_2$(ratio) equal to zero. For the remainder of this discussion, $\log_2$(ratio) will be used to represent expression levels.

Normalization

Typically, the first transformation applied to expression data, referred to as normalization, adjusts the individual hybridiza-

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA (e-mail: johnq@tigr.org).

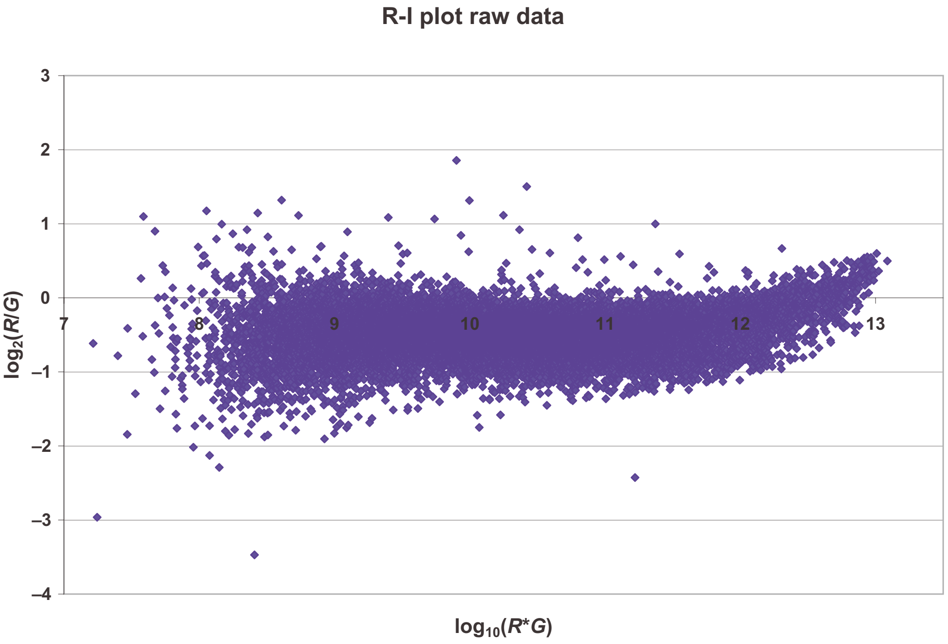

Fig. 1 An R-I plot displays the $\log_2(R_i/G_i)$ ratio for each element on the array as a function of the $\log_{10}(R_i{*}G_i)$ product intensities and can reveal systematic intensity-dependent effects in the measured $\log_2$(ratio) values. Data shown here are for a 27,648-element spotted mouse cDNA array.

tion intensities to balance them appropriately so that meaningful biological comparisons can be made. There are a number of reasons why data must be normalized, including unequal quantities of starting RNA, differences in labeling or detection efficiencies between the fluorescent dyes used, and systematic biases in the measured expression levels. Conceptually, normalization is similar to adjusting expression levels measured by northern analysis or quantitative reverse transcription PCR (RT–PCR) relative to the expression of one or more reference genes whose levels are assumed to be constant between samples.

There are many approaches to normalizing expression levels. Some, such as total intensity normalization, are based on simple assumptions. Here, let us assume that we are starting with equal quantities of RNA for the two samples we are going to compare. Given that there are millions of individual RNA molecules in each sample, we will assume that the average mass of each molecule is approximately the same, and that, consequently, the number of molecules in each sample is also the same. Second, let us assume that the arrayed elements represent a random sampling of the genes in the organism. This point is important because we will also assume that the arrayed elements randomly interrogate the two RNA samples. If the arrayed genes are selected to represent only those we know will change, then we will likely over- or under-sample the genes in one of the biological samples being compared. If the array contains a large enough assortment of random genes, we do not expect to see such bias. This is because for a finite RNA sample, when representation of one RNA species increases, representation of other species must decrease. Consequently, approximately the same number of labeled molecules from each sample should hybridize to the arrays and, therefore, the total hybridization intensities summed over all elements in the arrays should be the same for each sample.

Using this approach, a normalization factor is calculated by summing the measured intensities in both channels

$$N_{total} = \frac{\sum_{i=1}^{Narray} R_i}{\sum_{i=1}^{Narray} G_i},$$

where G_i and R_i are the measured intensities for the ith array element (for example, the green and red intensities in a two-color microarray assay) and N_{array} is the total number of elements represented in the microarray. One or both intensities are appropriately scaled, for example,

$$G'_k = N_{total} G_k \text{ and } R'_k = R_k,$$

so that the normalized expression ratio for each element becomes

$$T_i = \frac{R_i}{G_i} = \frac{1}{N_{total}} \frac{R_i}{G_i},$$

which adjusts each ratio such that the mean ratio is equal to 1. This process is equivalent to subtracting a constant from the logarithm of the expression ratio,

$$\log_2(T'_i) = \log_2(T_i) - \log_2(N_{total}),$$

which results in a mean $\log_2$(ratio) equal to zero.

There are many variations on this type of normalization, including scaling the individual intensities so that the mean or median intensities are the same within a single array or across all arrays, or using a selected subset of the arrayed genes rather than the entire collection.

Lowess normalization

In addition to total intensity normalization described above, there are a number of alternative approaches to normalizing expression ratios, including linear regression analysis[1], log centering, rank invariant methods[2] and Chen's ratio statistics[3], among others. However, none of these approaches takes into account systematic biases that may appear in the data. Several reports have indicated that the $\log_2$(ratio) values can have a systematic dependence on intensity[4,5], which most commonly appears as a deviation from zero for low-intensity spots. Locally weighted linear regression (lowess)[6] analysis has been proposed[4,5] as a normalization method that can remove such intensity-dependent effects in the $\log_2$(ratio) values.

The easiest way to visualize intensity-dependent effects, and the starting point for the lowess analysis described here, is to plot the measured $\log_2(R_i/G_i)$ for each element on the array as a function of the $\log_{10}(R_i{*}G_i)$ product intensities. This 'R-I' (for ratio-intensity) plot can reveal intensity-specific artifacts in the $\log_2$(ratio) measurements (Fig. 1).

Lowess detects systematic deviations in the R-I plot and corrects them by carrying out a local weighted linear regression as a function of the $\log_{10}$(intensity) and subtracting the calculated best-fit average $\log_2$(ratio) from the experimentally observed ratio for each data point. Lowess uses a weight function that de-emphasies the contributions of data from array elements that are far (on the R-I plot) from each point.

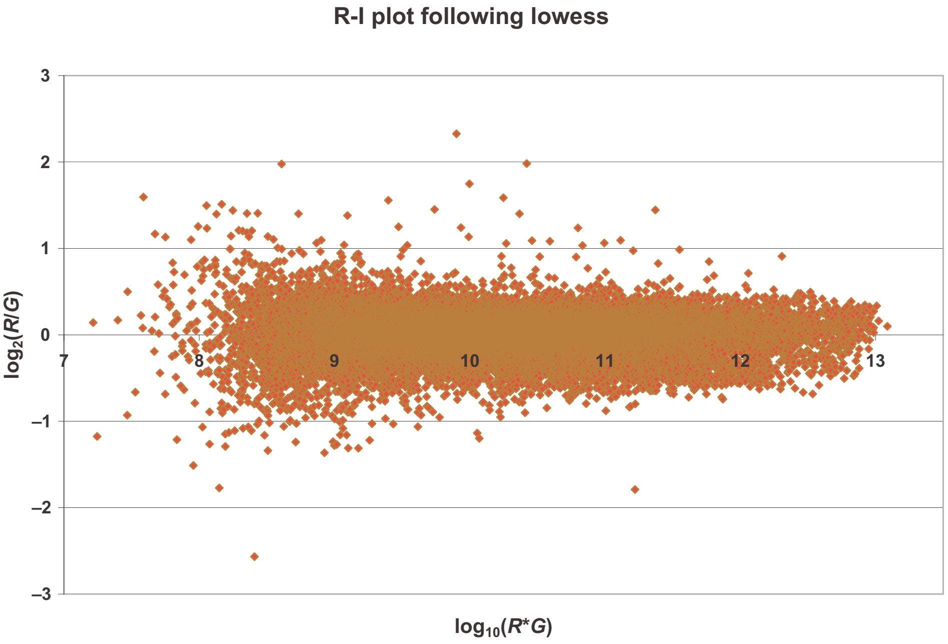

Fig. 2 Application of local (pen group) lowess can correct for both systematic variation as a function of intensity and spatial variation between spotting pens on a DNA microarray. Here, the data from Fig. 1 has been adjusted using a lowess in the TIGR MIDAS software (with a tri-cube weight function and a 0.33 smooth parameter) available from http://www.tigr.org/software.

If we set $x_i = \log_{10}(R_i{*}G_i)$ and $y_i = \log_2(R_i/G_i)$, lowess first estimates $y(x_k)$, the dependence of the $\log_2$(ratio) on the $\log_{10}$(intensity), and then uses this function, point by point, to correct the measured $\log_2$(ratio) values so that

$$\log_2(T_i')=\log_2(T_i)-y(x_i)=\log_2(T_i)-\log_2(2^{y(x_i)}),$$

or equivalently,

$$\log_2(T_i')=\log_2\left(T_i * \frac{1}{2^{y(xi)}}\right)=\log_2\left(\frac{R_i}{G_i} * \frac{1}{2^{y(xi)}}\right).$$

As with the other normalization methods, we can make this equation equivalent to a transformation on the intensities, where

$$G_i'= G_i * 2^{y(x_i)} \text{ and } R_i'= R_i.$$

The results of applying such a lowess correction can be seen in Fig. 2.

Global versus local normalization. Most normalization algorithms, including lowess, can be applied either globally (to the entire data set) or locally (to some physical subset of the data). For spotted arrays, local normalization is often applied to each group of array elements deposited by a single spotting pen (sometimes referred to as a 'pen group' or 'subgrid'). Local normalization has the advantage that it can help correct for systematic spatial variation in the array, including inconsistencies among the spotting pens used to make the array, variability in the slide surface, and slight local differences in hybridization conditions across the array. When a particular normalization algorithm is applied locally, all the conditions and assumptions that underlie the validity of the approach must be satisfied. For example, the elements in any pen group should not be preferentially selected to represent differentially expressed genes, and a sufficiently large number of elements should be included in each pen group for the approach to be valid.

Variance regularization. Whereas normalization adjusts the mean of the $\log_2$(ratio) measurements, stochastic processes can cause the variance of the measured log2(ratio) values to differ from one region of an array to another or between arrays. One approach to dealing with this problem is to adjust the $\log_2$(ratio) measures so that the variance is the same[4,7]. If we consider a single array with distinct subgrids for which we have carried out local normalization, then what we are seeking is a factor for each subgrid that we can use to scale all of the measurements within that subgrid.

An appropriate scaling factor is the variance for a particular subgrid divided by the geometric mean of the variances for all subgrids. If we assume that each subgrid has M elements, because we have already adjusted the mean of the $\log_2$(ratio) values in each subgrid to be zero, their variance in the nth subgrid is

$$\sigma_n^2=\sum_{j=1}^{M}\left[\log_2(T_j)\right]^2,$$

where the summation runs over all the elements in that subgrid. If the number of subgrids in the array is N_{grids}, then the appropriate scaling factor for the elements of the kth subgrid on the array is

$$a_k=\frac{\sigma_k^2}{\left[\prod_{n=1}^{N_{grids}}\sigma_n^2\right]^{1/N_{grids}}}.$$

We then scale all of the elements within the kth subgrid by dividing by the same value a_k computed for that subgrid,

$$\log_2(T_i)=\frac{\log_2(T_i)}{a_k}.$$

This is equivalent to taking the a_kth root of the individual intensities in the kth subgrid,

$$G_i'=\left[G_i\right]^{1/a_k} \text{ and } R_i'=\left[R_i\right]^{1/a_k}.$$

It should be noted that other variance regularization factors have been suggested[4] and that, obviously, a similar process can be used to regularize variances between normalized arrays.

Intensity-based filtering of array elements

If one examines several representative R-I plots, it becomes obvious that the variability in the measured $\log_2$(ratio) values increases as the measured hybridization intensity decreases. This is not surprising, as relative error increases at lower intensities, where the signal approaches background. A commonly used approach to address this problem is to use only array elements with intensities that are statistically significantly different from background. If we measure the average local background near each array element and its standard deviation, we would expect at 95.5% confidence that good elements would have intensities more than two standard deviations above background. By keeping only array elements that are confidently above background,

$$G_i^{spot}>2*\sigma(G_i^{background}) \text{ and } R_i^{spot}>2*\sigma(R_i^{background}),$$

we can increase the reliability of measurements. Other approaches include the use of absolute lower thresholds for acceptable array

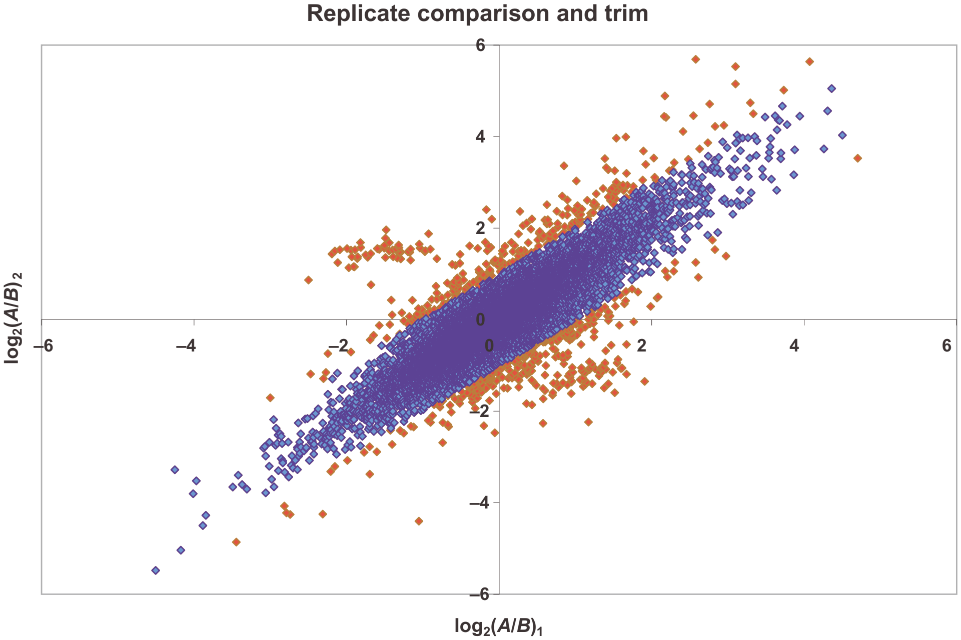

Fig. 3 The use of replicates can help eliminate questionable or inconsistent data from further analysis. Here, the lowess-adjusted $\log_2(A_i/B_i)$ values for two independent replicates are plotted against each other element by element for hybridizations to a 32,448-element human array. Outliers in the original data (in red) are excluded from the remainder of the data (blue) selected on the basis of a two-standard-deviation cut on the replicates.

elements (sometimes referred to as 'floors') or percentage-based cut-offs in which some fixed fraction of elements is discarded.

A different problem can occur at the high end of the intensity spectrum, where the array elements saturate the detector used to measure fluorescence intensity. Once the intensity approaches its maximum value (typically 2^{16}–1=65,535 per pixel for a 16-bit scanner), comparisons are no longer meaningful, as the array elements become saturated, and intensity measurements cannot go higher. Again, there are a variety of approaches to dealing with this problem as well, including eliminating saturated pixels in the image-processing step or setting a maximum acceptable value (often referred to as a 'ceiling') for each array element.

Replicate filtering

Replication is essential for identifying and reducing the variation in any experimental assay, and microarrays are no exception. Biological replicates use RNA independently derived from distinct biological sources and provide both a measure of the natural biological variability in the system under study, as well as any random variation used in sample preparation. Technical replicates provide information on the natural and systematic variability that occurs in performing the assay. Technical replicates include replicated elements within a single array, multiple independent elements for a particular gene within an array (such as independent cDNAs or oligos for a particular gene), or replicated hybridizations for a particular sample. The particular approach used will depend on the experimental design and the particular study underway (see also the review by G. Churchill, pages 490–495, this issue)[8].

To illustrate the usefulness of technical replicates, consider their use in identifying and eliminating low-quality or questionable array elements. One widely used technical replication in two-color spotted array analysis is dye-reversal or flip-dye analysis[8], which consists of duplicating labeling and hybridization by swapping the fluorescent dyes used for each RNA sample. This process may help to compensate for any biases that may occur during labeling or hybridization; for example, if some genes preferentially label with the red or green dye. Let us assume we have two samples, A and B. In the first hybridization, we label A with our red dye and B with our green dye and reverse the dye labeling in the second, so that the ratios for our measurements can be defined, respectively, as

$$T_{1i} = \frac{R_{1i}}{G_{1i}} = \frac{A_{1i}}{B_{1i}} \text{ and } T_{2i} = \frac{R_{2i}}{G_{2i}} = \frac{B_{2i}}{A_{2i}}\ .$$

As we are making two comparisons between identical samples, we expect

$$(T_{1i} * T_{2i}) = \left(\frac{A_{1i}}{B_{1i}} * \frac{B_{2i}}{A_{2i}}\right) = 1,$$

or equivalently,

$$\log_2(T_{1i} * T_{2i}) = \log_2\left(\frac{A_{1i}}{B_{1i}} * \frac{B_{2i}}{A_{2i}}\right) = 0.$$

However, we know that experimental variation will lead to a distribution of the measured values for the log of the product ratios, $\log_2(T_{1i}{*}T_{2i})$. For this distribution, we can calculate the mean and standard deviation. One would expect the consistent array elements to have a value for $\log_2(T_{1i}{*}T_{2i})$ 'close' to zero and inconsistent measurements to have a value 'far' from zero. Depending on how stringent we want to be, we can choose to keep and use array element data for which $\log_2(T_{1i}{*}T_{2i})$ is within a certain number of standard deviations of the mean. Although one cannot determine *a priori* which of the replicates is likely to be in error, visual inspection may allow the 'bad' element to be identified and removed before further analysis. Alternatively, and more practically for large experiments, one can simply eliminate questionable elements from further consideration. The application of such a replicate trim is shown in Fig. 3. Obviously, similar approaches can be used to filter data from other replication strategies, including replicates within a single array or measurements from replicated single-color arrays.

Averaging over replicates

To reduce the complexity of the data set, we may average over the replicate measures. If, as before, we consider replicate measures of two samples, A and B, we would want to adjust the $\log_2$(ratio) measures for each gene so that the transformed values are equal, or, for the ith array element,

$$\log_2\left(\frac{A_{1i}}{B_{1i}}\right) + c_i = \log_2\left(\frac{A_{2i}}{B_{2i}}\right) - c_i.$$

This equation can be easily solved to yield a value for the constant c_i that is used to correct each array element,

$$c_i = \frac{1}{2}\log_2\left(\frac{A_{2i}}{B_{2i}}\frac{B_{1i}}{A_{1i}}\right) = \log_2\left(\sqrt{\frac{A_{2i}}{B_{2i}}\frac{B_{1i}}{A_{1i}}}\right).$$

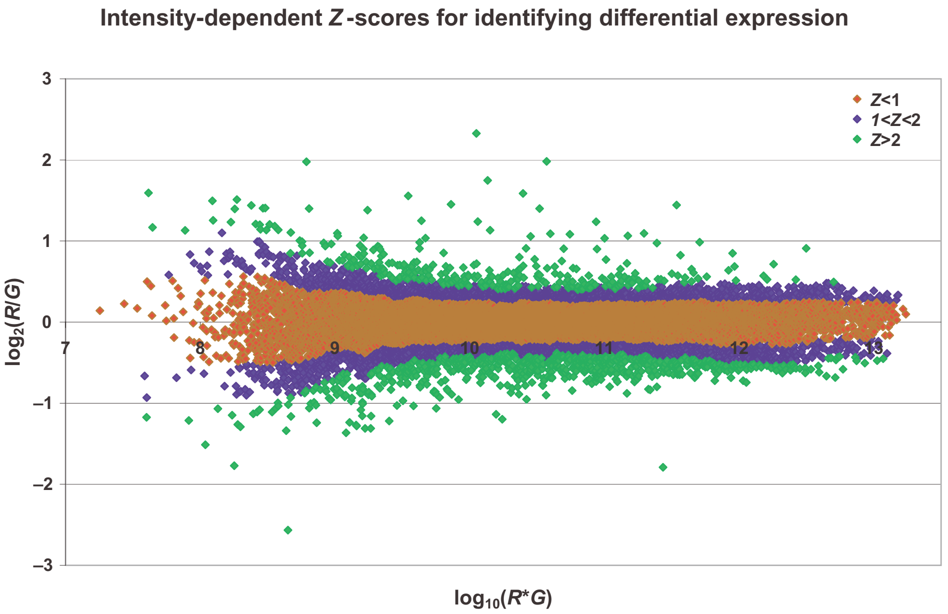

Fig. 4 Local variation as a function of intensity can be used to identify differentially expressed genes by calculating an intensity-dependent Z-score. In this R-I plot, array elements are color-coded depending on whether they are less than one standard deviation from the mean (blue), between one and two standard deviations (red), or more than two standard deviations from the mean (green).

If we use this equation to average our measurements, the result is equivalent to taking the geometric mean, or

$$\log_2\left(\frac{\bar{A}_i}{\bar{B}_i}\right) = \log_2\left(\sqrt{\frac{A_{2i}}{B_{2i}}\frac{A_{1i}}{B_{1i}}}\right),$$

where the average measurements for expression in each sample is given by

$$\bar{A}_i = \sqrt{A_{1i}A_{2i}} \text{ and } \bar{B}_i = \sqrt{B_{1i}B_{2i}}.$$

The adjusted average measures

$$\bar{A}_i \text{ and } \bar{B}_i,$$

and for each gene can then be used to carry out further analyses. For example, one can create an R-I plot, with

$$\log_2(\bar{A}_i / \bar{B}_i)$$

plotted as a function of the

$$\log_{10}(\bar{A}_i * \bar{B}_i)$$

product intensities for each arrayed element, or for any other application. This procedure can obviously be extended to averaging over n replicates by taking the nth root.

Propagation of errors

One advantage of having replicates is that they can be used to estimate uncertainty in derived quantities based on the measured or estimated uncertainties in the measured quantities. The theory of error propagation[9] tells us that, in general, if we have a function $f(A,B,C,\ldots)$, with error estimates σ for each of the input variables, $A,B,C,\ldots$, then the squared error in f is

$$\sigma_f^2 = \sigma_A^2\left[\frac{\partial f}{\partial A}\right]^2 + \sigma_B^2\left[\frac{\partial f}{\partial B}\right]^2 + \ldots,$$

where

$$\frac{\partial f}{\partial A}$$

is the partial derivative of f with respect to A. For microarrays, the function with which we are most often concerned is the $\log_2$(ratio),

$$f = \log_2\left(\frac{A_1}{B_1}\right) = \log_2(A_i) – \log_2(B_i).$$

Here, we can take advantage of the fact that logarithms can easily be changed between bases and that we know the derivative of the natural logarithm,

$$\log_2(x) = \frac{\ln(x)}{\ln(2)} \text{ and } \frac{\partial}{\partial x}[\ln(x)] = \frac{1}{x},$$

to derive the following equation

$$\frac{\partial}{\partial x}[\log_2(x)] = \frac{\partial}{\partial x}\left[\frac{\ln(x)}{\ln(2)}\right] = \left[\frac{1}{\ln(2)}\right]\frac{\partial}{\partial x}[\ln(x)] = \left[\frac{1}{\ln(2)}\right]\left[\frac{1}{x}\right] = \frac{1}{x\ln(2)}.$$

It then follows that the squared-error in our average $\log_2$(ratio) is

$$\sigma_f^2 = \sigma_A^2\left[\frac{\partial f}{\partial A}\right]^2 + \sigma_B^2\left[\frac{\partial f}{\partial B}\right]^2 = \sigma_A^2\left[\frac{1}{A\ln(2)}\right]^2 + \sigma_B^2\left[\frac{1}{B\ln(2)}\right]^2$$

so that the errors in the expression measures for A and B can be used to estimate the error in the $\log_2$(ratio).

Identifying differentially expressed genes

Regardless of the experiment performed, one outcome that is invariably of interest is the identification of genes that are differentially expressed between one or more pairs of samples in the data set. Even if data-mining analysis is going to be done using, for example, one or more of the widely used clustering methods[10–12], it is still extremely useful to reduce the data set to those genes that are most variable between samples. In many early microarray analyses, a fixed fold-change cut-off (generally two-fold) was used to identify the genes exhibiting the most significant variation. A slightly more sophisticated approach involves calculating the mean and standard deviation of the distribution of $\log_2$(ratio) values and defining a global fold-change difference and confidence; this is essentially equivalent to using a Z-score for the data set. In an R-I plot, such criteria would be represented as parallel horizontal lines; genes outside of those lines would be called differentially expressed.

Analysis of the R-I plot suggests, however, that this approach may not accurately reflect the inherent structure in the data. At low intensities, where the data are much more variable, one might misidentify genes as being differentially expressed, while at higher intensities, genes that are significantly expressed might not be identified. An alternative approach would be to use the local structure of the data set to define differential expression. Using a sliding window, one can calculate the mean and standard deviation within a window surrounding each data point, and define an intensity-dependent Z-score threshold to identify differential expression[5], where Z simply measures the number of standard deviations a particular data point is from the mean. If

$$\sigma^{local}_{\log_2(T_i)}$$

is the calculated standard deviation in a region of the R-I plot surrounding the $\log_2$(ratio) for a particular array element i, then

$$Z_i^{local} = \frac{\log_2(T_i)}{\sigma^{local}_{\log_2(T_i)}}.$$

With this definition, differentially expressed genes at the 95% confidence level would be those with a value of

$$|Z_i^{local}| > 1.96,$$

or, equivalently, those more than 1.96 standard deviations from the local mean. At higher intensities, this allows smaller, yet still significant, changes to be identified, while applying more stringent criteria at intensities where the data are naturally more variable. An example of the application of a Z-score selection is shown in Fig. 4.

Having identified differentially expressed genes for each pair of hybridized samples, one can then further filter the entire experimental data set to select for further analysis only those genes that are differentially expressed in a subset of the experimental samples, such as disease or normal tissues, or using any other criteria that make sense for the experimental design used. Another alternative would be to use the ANOVA techniques summarized by Churchill on pages 490–495 of this supplement[8] to select significantly differentially expressed genes. In either case, the resulting reduced data set can then be used for further data mining and analysis.

Looking ahead

There is a wide range of additional transformations that can be applied to expression data, and we have presented only a small sample of the available techniques that can be used. For example, there are specific transformations that have been developed for analysis of data from particular platforms[13,14], and sophisticated methods have been proposed for development of error models based on analysis of repeated hybridizations[14,15]. Regardless of the sophistication of the analysis, nothing can compensate for poor-quality data. The single most important data-analysis technique is the collection of the highest-quality data possible. The starting point for effective data collection and analysis is a good experimental design with sufficient replication to ensure that both the experimental and biological variation can be identified and estimated. A second important element in generating expression data is optimization and standardization of the experimental protocol. Like northern blots and RT–PCR, microarrays directly assay relative RNA levels and use these to infer gene expression. Therefore, at every step in the process, from sample collection through RNA isolation, array preparation, sample labeling, hybridization, data collection and analysis, every possible effort must be made to minimize variation.

Defining objective criteria for the quality of a DNA microarray assay remains an open problem, but one that clearly needs to be addressed as microarray assays become more widespread. One could easily write another review article on elements that contribute to good-quality arrays and which could be used, in part, to define such a quality standard. The challenge is to use these often qualitative objectives to define one or more appropriate quantitative measure. The Normalization Working Group of the Microarray Gene Expression Data (MGED) organization (http://www.mged.org) is attempting to define such a standard, and community input and participation is welcome (see commentary by C. Stoeckert, pages 469–473, this issue)[16].

Finally, as standards for publication of microarray experiments evolve, reporting data transformations is becoming as important as disclosing laboratory methods. Without an accurate description of either of these, it would be difficult, if not impossible, for the results derived from any study to be replicated.

Acknowledgments

The work presented here evolved from looking at a large body of data and would have been much less useful without the contributions of Norman H. Lee, Renae L. Malek, Priti Hegde, Ivana Yang, Shuibang Wang, Yonghong Wang, Simon Kwong, Heenam Kim, Wei Liang, Vasily Sharov, John Braisted, Alex Saeed, Joseph White, Jerry Li, Renee Gaspard, Erik Snesrud, Yan Yu, Emily Chen, Jeremy Hasseman, Bryan Frank, Lara Linford, Linda Moy, Tara Vantoai, Gary Churchill and Roger Bumgarner. J.Q. is supported by grants from the US National Science Foundation, the National Heart, Lung, and Blood Institute, and the National Cancer Institute. The MIDAS software system used for the normalization and data filtering presented here is freely available as either executable or source code from http://www.tigr.org/ software, along with the MADAM data-management system, the Spotfinder image-processing software, and the MeV clustering and data-mining tool.

1. Chatterjee, S. & Price, B. *Regression Analysis by Example* (John Wiley & Sons, New York, 1991).
2. Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. & Wong, W.H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29**, 2549–2557 (2001).
3. Chen, Y., Dougherty, E.R. & Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* **2**, 364–374 (1997).
4. Yang, Y.H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
5. Yang, I.V. *et al.* Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.* **3**, research0062.1–0062.12 (2002).
6. Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Stat. Assoc.* **74**, 829–836 (1979).
7. Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18,** S96–S104 (2002).
8. Churchill, G.A. Fundamentals of experimental design for cDNA microarrays. *Nature Genet.* **32**, 490–495 (2002).
9. Bevington, P.R. & Robinson, D.K. *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, New York, 1991).
10. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
11. Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. & Somogy, R. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA* **95**, 334–339 (1998).
12. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
13. Li, C. & Wong, W. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA* **98**, 31–36 (2001).
14. Ideker, T., Thorsson, V., Siegel, A.F. & Hood, L.E. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.* **7**, 805–817 (2001).
15. Rocke, D. & Durbin, B. A model for measurement error for gene expression arrays. *J. Comput. Biol.* **8**, 557–569 (2001).
16. Stoeckert, C.J., Causton, H.C. & Ball, C.A. Microarray databases: standards and ontologies. *Nature Genet.* **32**, 469–473 (2002).