BMC Biotechnology

BioMed Central

Methodology article

Open Access

The limits of log-ratios

Vasily Sharov[1], Ka Yin Kwong[1], Bryan Frank[1], Emily Chen[1], Jeremy Hasseman[1], Renee Gaspard[1], Yan Yu[1], Ivana Yang[1] and John Quackenbush*[1,2,3]

Address: [1]The Institute for Genomic Research, Rockville, MD, USA, [2]Department of Biochemistry, The George Washington University School of Medicine, Washington, DC, USA and [3]Department of Biostatistics, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD, USA

Email: Vasily Sharov - vsharov@tigr.org; Ka Yin Kwong - kkwong@tigr.org; Bryan Frank - bfrank@tigr.org; Emily Chen - echen@tigr.org; Jeremy Hasseman - hasseman@tigr.org; Renee Gaspard - rgaspard@tigr.org; Yan Yu - yyu@tigr.org; Ivana Yang - iyang@tigr.org; John Quackenbush* - johnq@tigr.org

* Corresponding author

Published: 08 March 2004

BMC Biotechnology 2004, **4**:3

Received: 15 November 2003
Accepted: 08 March 2004

This article is available from: http://www.biomedcentral.com/1472-6750/4/3

Abstract

Background: DNA microarray assays typically compare two biological samples and present the results of those comparisons gene-by-gene as the logarithm base two of the ratio of the measured expression levels for the two samples.

Results: Because of the fixed dynamic range of fluorescence and other detection systems, there is a limit to the range of comparisons that can be made using any array technology, and this must be taken into account when interpreting the results of any such analysis.

Conclusions: The dynamic range of microarray data collection systems results in limits in the comparative analyses that can be derived from such measurements and suggests that optimal results can be obtained by making measurements that avoid the boundaries of that dynamic range.

Background

DNA microarray analysis has become one of the most widely used techniques in modern molecular genetics, and the laboratory protocols that have developed in recent years have led to increasingly robust assays. The application of microarray technologies affords great opportunities for exploring patterns of gene expression and allows users to begin investigating problems ranging from deducing biological pathways to classifying patient populations.

As with all assays, the starting point for developing a microarray study is planning the comparisons that will be made, and the simplest experimental designs are based on the comparative analysis of two classes of samples, either using a series of paired case-control comparisons, or comparisons to a common reference sample, although other approaches have been described. But the fundamental question addressed using arrays is generally a comparison between paired samples to find genes that are significantly different in their patterns of expression. For the sake of the analysis presented here, we will focus on direct pair-wise comparisons between samples using spotted DNA arrays conducted as dual-labeled co-hybridization assays. However, it must be noted that the results we present here will impact other analyses including inferred relative changes derived by comparisons to a reference sample, through more complex loop designs, or from comparisons between single-color assays such as those which are commonly performed using the Affymetrix GeneChip™ or filter array platforms.

Results and Discussion

Measuring log-ratios on microarrays

Microarray experiments generally measure relative expression levels between biological samples. However, there is a fundamental limit to the changes that can be measured on an array and understanding that that these limits exist is important for analyzing microarray experiments. This observation depends fundamentally on the manner in which most microarray scanners work. Following hybridization of spectrally distinguishable labeled targets to the arrayed probes on a microarray, the surface of the slide is generally interrogated using one or more lasers, each tuned to excite a particular fluorescent label. The fluorescent light emitted from the surface is collected through an optical system, generally spectrally separated, and focused on a photon detector, usually a photomultiplier tube (PMT). PMTs have a glass photocathode window coated by one or more alkali metals that has a high probability of converting an incoming photon to an electron. The electron emitted from the window is attracted to an alkali metal coated electrode which is maintained at a positive charge. When the initial electron strikes the electrode, it normally releases a number of additional electrons. These are attracted to a series of coated electrodes, each maintained at a slightly higher voltage than the previous, in effect multiplying the number of electrons released at each subsequent electrode. After a series of these amplification steps, the electrons are collected by a final electrode and the output current is measured. This output current depends on the intensity of the light (*i.e.* the number of photons) and the total voltage maintained across the PMT – a higher voltage accelerates electrons more in each step, producing a greater final current. It should be noted that this process is also stochastic, so that each photon produces a number of electrons which can be modeled as a Gaussian distribution with mean μ and standard deviation σ. It should be noted that as the light intensity increases, the number of photons increases and this has an effect on the distribution, with N photons producing approximately $N\mu$ final electrons with a standard deviation of $\sigma/\sqrt{N}$. This explains, in part, the reason why the variation in signal intensity, and consequently derived measurements such as log-ratios are more uncertain for genes expressed at lower levels. Finally, the signal from the PMT is converted to a digital signal using an analog-to-digital converter (ADC). Typical array scanners use 16-bit ADCs, giving the instruments an output range of 0 to 65535 (2^{16}-1) relative fluorescence units (RFUs) for each pixel. The reported intensity values for each spot on the array varies between research groups and software used for image processing. Common measures of expression include background subtracted mean or median pixel values measures for each arrayed gene. For the purposes of the analysis presented here, we will use the background-subtracted mean pixel values reported by the TIGR Spotfinder image analysis software [1].

Microarray assays are often used to compare expression levels between paired samples and for a variety of reasons, these comparisons are typically expressed for each gene as the logarithm base 2 of the ratio of the (background subtracted) fluorescent signals measured from each labeled sample [$\log_2(R/G)$]; we refer to these as log-ratios. Because the fluorescent dyes used in most microarray assays have slightly different efficiencies for light emission, the detection efficiencies of the phototubes has some wavelength dependence and hence differ for the different dyes, and because the PMTs exhibit nonlinearities at high and low intensities, the log-ratios measured often exhibit some systematic, intensity-dependent variation. This systematic error is most easily visualized using a Ratio-Intensity (RI) plot ([2,3]; also called an MA plot by Speed and colleagues) in which the log-ratio for each spot is plotted as a function of one-half the logarithm of the product of the measured intensity $\left[\frac{1}{2}\log_2\left(R * G\right)\right]$, which is equivalent to the logarithm of the geometric mean of the intensity for that gene, a measure of the relative expression level of a particular gene. The shape of the distribution one observes in an R-I plot depends in a fundamental way on the experimental design one chooses as that defines the comparisons that are made. For closely related samples where one expects gene expression to be highly similar, the distribution of log-ratio values is broad at lower intensities, reflecting the greater relative uncertainty as one approaches the detection limits in one or both channels, while it narrows at higher expression levels (Figure 1A,1B); for biologically diverse samples the R-I plot can present a very different profile (Figure 1C,1D,1E,1F).

The R-I plot can also reveal some of the limitations of using log-ratios as a measure of expression. As described previously, the 16-bit ADCs in microarray scanners limit the maximum intensity that can be measured in both red and green channels on an array such that both $\log_2(R)$ and $\log_2(G)$ values range independently between a minimum of 0 and a maximum of 16. One can visualize this as a square box in the a plot of $\log_2(R)$ versus $\log_2(G)$, or as a diamond-shaped area in an R-I plot (Figure 2A). This relationship is due to the fact that the R-I plot is essentially a 45° ($\pi/4$) rotation (and slight rescaling) of the log-intensity plot, where the square represents the limits defined by each of the two independent fluorescence measurements (Figure 2B).

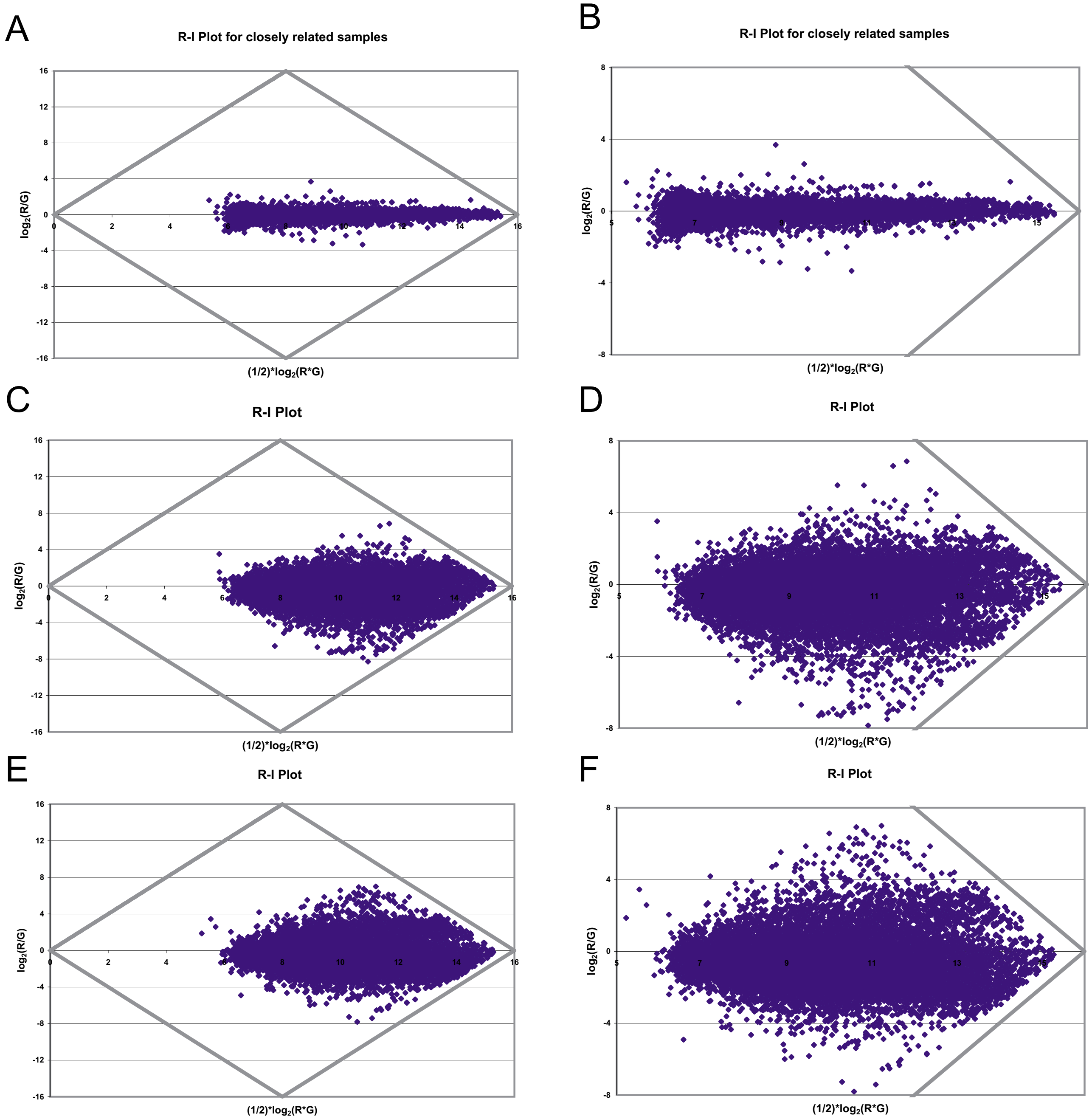

Figure 1

R-I plots for microarray expression data exhibit the limits of log-ratio measurements obtained on arrays. As the measured intensity on the arrays approaches its upper and lower limits, the dynamic range for accurately estimating fold-change measurements is also limited. Shown here are R-I plots for three different data sets showing the entire range (A,C,E) and a close-up of the upper end of the end of the effective range for array measurements (respectively B,D,F). The diamond-shape delimits the range of measurements obtainable on microarrays.

Most microarray image analysis software performs a background subtraction and uses other methods to avoid saturation of pixels, the reported fluorescence signals normally do not reach the absolute limit of detection. The

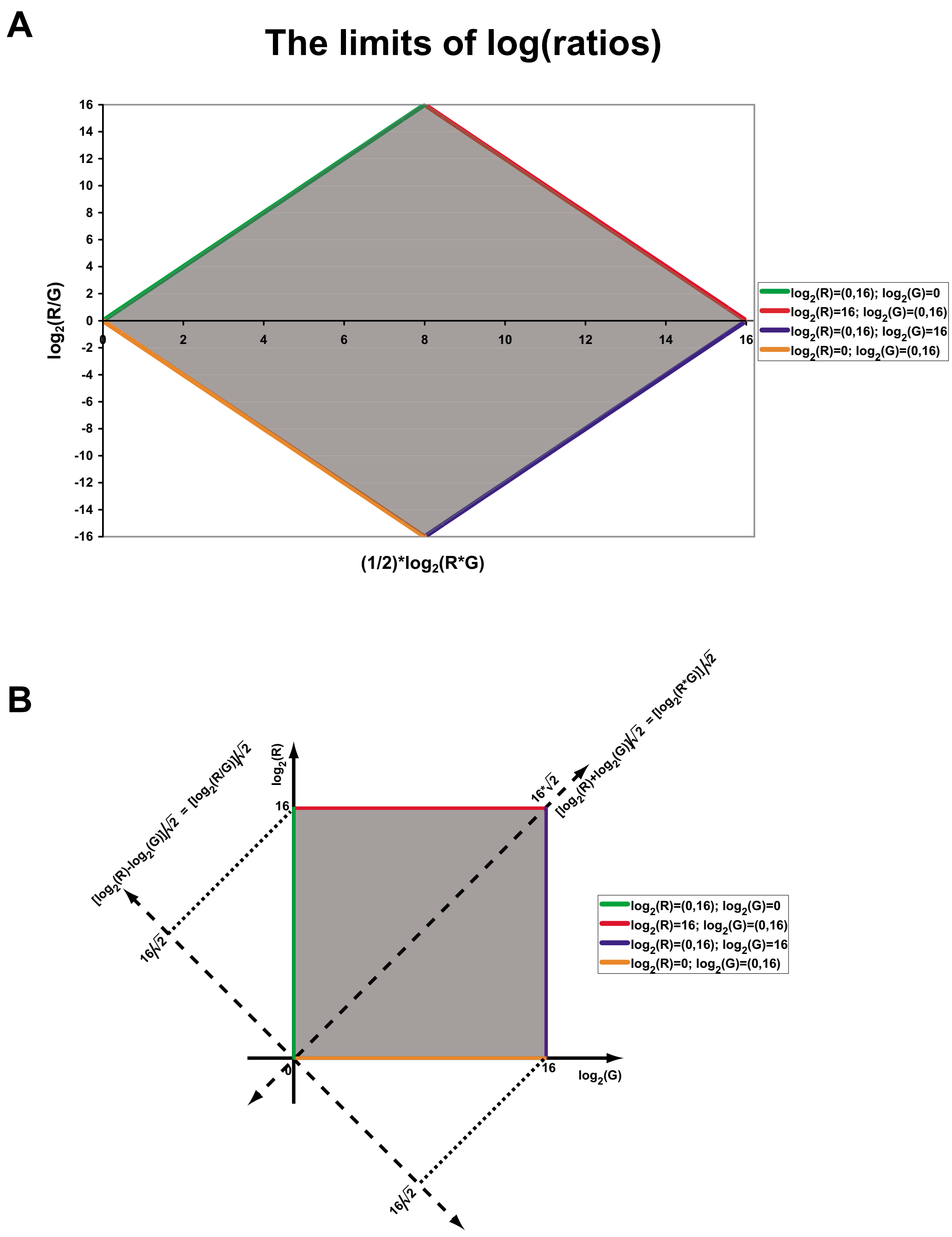

Figure 2
The limitation on the dynamic range of log-ratio measurements, (A) shown here in the diamond-shaped gray-shaded box between the colored lines on an R-I plot, reflects the limited range of values that can be obtained from existing microarray technology which typically employ 16-bit array scanners that allow each channel on the arrays to produce measurements ranging in $\log_2$ values from 0 through 16. (B) The diamond area represents a rotation of the original axes, $x = \log_2(G)$ and $y = \log_2(R)$ to new axes $x' = [\log_2(R) + \log_2(G)]/\sqrt{2}$ and $y' = [\log_2(R) - \log_2(G)]/\sqrt{2}$, followed by a simple rescaling to $x'' = [\log_2(R) + \log_2(G)]/2$ and $y'' = [\log_2(R) - \log_2(G)]$.

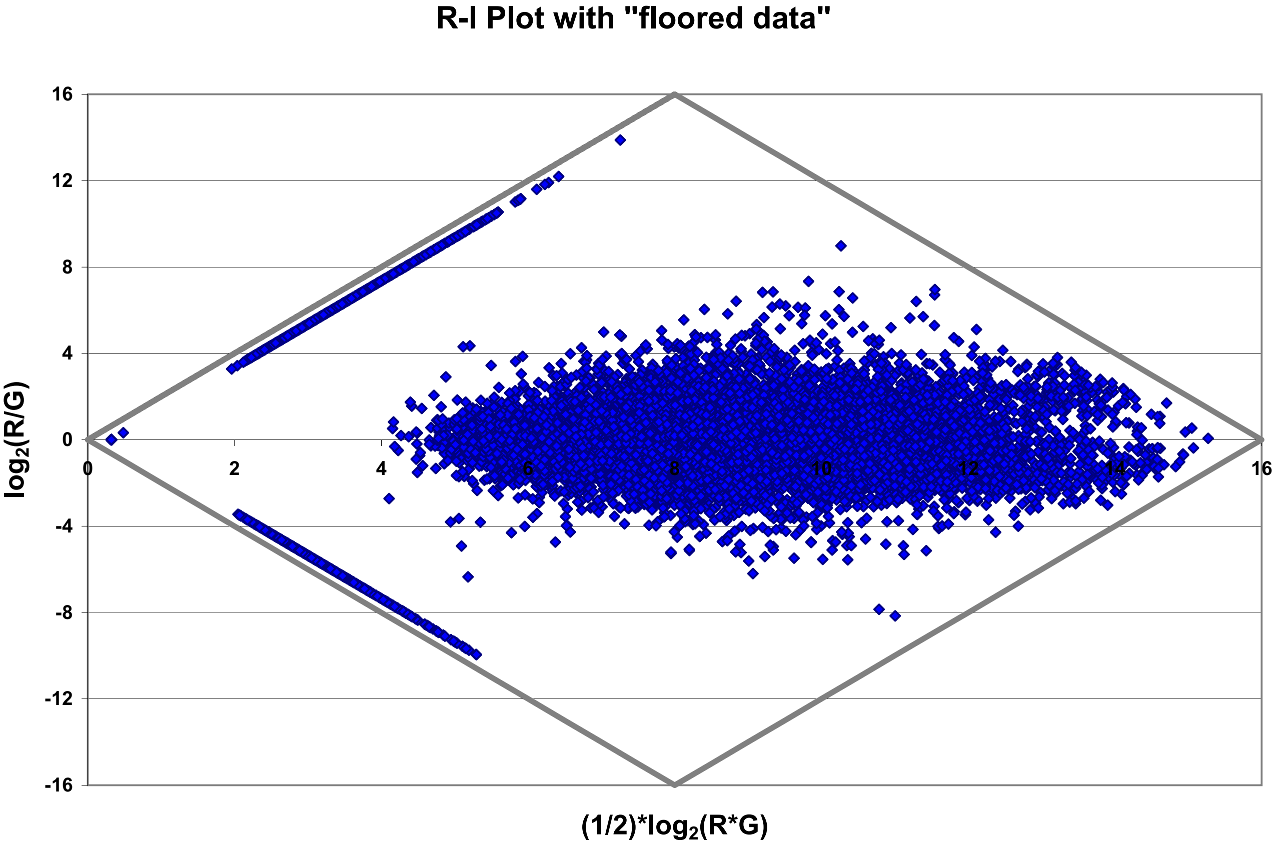

Figure 3
A common practice for low-intensity elements on the array is to set a "floor" representing a minimum intensity that is reported. This eliminates "undefined" log-ratio values that come from reported zeros, but produces "whiskers" in the R-I plot.

background-subtracted data we use for analysis exhibit that effect in hybridization assays where the fluorescence signal is particularly strong (Figure 1B,1D,1F). Similar effects can be seen as the signal intensity decreases toward the lower limit, where discrete integer values assigned to gene expressed at low levels appear as diagonal "whiskers" in the R-I plot (Figure 3); this often arises as a result of setting expression values below some threshold to a minimal value, a process referred to as "flooring."

It is important to note that this effect limits the dynamic range of "fold-change" (equivalent to the log-ratio) measurements on arrays, particularly as the measured intensities approach either the minimum or maximum detectable levels accessible on a particular array scanner. Furthermore, it is important to note that these limits are not unique to dual-color detection techniques. Comparisons made using single color microarrays are also limited by the dynamic range of the individual measurements and fold-change estimates in comparisons demonstrate exactly the same type of artifact.

Conclusions

This simple analysis presented here suggests a possible limitation on the use of fold-change measurements derived from microarrays and argues for the use of R-I plots as a means of detecting possible deviations from the dynamic range of the assay. Further, these results suggest that rather than try to maximize signal on the fluorescent images from the array, a better approach would be to target background-subtracted fluorescent intensities to the middle of the range where the dynamic range for fold change measurements is maximized, or a $\left[\frac{1}{2}\log_2(R*G)\right]$ of 8. However, this corresponds to an average expression measurement of only 256 RFUs, which on most arrays is uncomfortably close to background. In practice, an average $\left[\frac{1}{2}\log_2(R*G)\right]$ of 10 to 12 (1024 to 4096) strikes a good balance between intensities that are too close to background and those that approach the limits of the dynamic range of the assay. While the raw images from these arrays may not provide as pretty a picture of the hybridization assay, they are more likely to provide useful data that can be validated.

Authors' contributions

VS and JQ are responsible for drafting the manuscript and producing the final version. KYK, BF, EC, JH, RG, YY, and IY contributed data and participated in its analysis. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by grants from the National Heart, Lung, Blood Institute (NIH 1 U01 HL66580-01 and NIH-1 R33 HL3712-01), the US National Cancer Institute (NIH-U01-CA8552-01A1), and the National Science Foundation NSF-DBI9975920 and NSF-DBI-0177281).

References

1. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J: **TM4: A Free, Open Source System for Microarray Data Management and Analysis.** *Biotechniques* 2003:374-378.
2. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biol* 2002, **3:**research0062.
3. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30:**e15.