[Frontiers in Bioscience E2, 171-186, January 1, 2010]

Cross phenotype normalization of microarray data

Jianhua Xuan[1], Yue Wang[1], Eric Hoffman[2], Robert Clarke[3]

[1]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA, [2]Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC 20010, USA, [3]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20007, USA

TABLE OF CONTENTS

1. ABSTRACT

Normalization is a prerequisite for almost all follow-up steps in microarray data analysis. Accurate normalization across different experiments and phenotypes assures a common base for comparative yet quantitative studies using gene expression data. In this paper, we report a comparison study of four normalization approaches, namely, linear regression (LR), Loess regression, invariant ranking (IR) and iterative nonlinear regression (INR) method, for gene expression. Among these four methods, LR and Loess regression methods use all available genes to estimate either a linear or nonlinear normalization function; while IR and INR methods feature some iterative processes to identify invariantly expressed genes (IEGs) for nonlinear normalization. We tested these normalization approaches on three real microarray data sets and evaluated their performance in terms of variance reduction and fold-change preservation. By comparison, we found that (1) LR method exhibits the worst performance in both variance reduction and fold-change preservation, and (2) INR method shows an improved performance in achieving low expression variance across replicates and excellent fold-change preservation for differently expressed genes.

2. INTRODUCTION

DNA microarray technology has enabled high-throughput measurements of tens of thousands of mRNA levels, providing us a powerful tool to investigate biochemical pathways and gene regulatory networks, to identify phenotype-specific biomarkers, to assess cellular response to drug compounds, and to classify disease states at molecular level. For example, recent studies in cancer research demonstrate that gene expression profiling can reveal distinct tumor subtypes not evident by traditional histopathological methods (1). Although it is optimistic to assume that gene expression data alone will be sufficient for the reconstruction of complete regulatory pathways, several recent studies successfully demonstrate the potential for inferring regulatory networks from gene expression data (3).

While high-throughput measurements of gene expression levels are likely to provide important information about cellular processes (e.g., revealing previously unrecognized patterns of gene regulation) and generate new hypotheses warranting further study, widespread use of microarray profiling methods is limited by the need for further technology developments, particularly computational bioinformatics tools not previously included by the instruments. Recently, much effort has been devoted to the development of high-level data analysis tools such as clustering (4-6), classification (2,7,8) and Bayesian network methods (3). As more and more computational tools are made available to researchers, it has become increasingly clear that the key issue in microarray data analysis is how to extract quality information about the biological system being studied.

As a first step in accurately exacting biological information, it is necessary to filter out experimental noise and correct for systematic errors confounding the raw data obtained by this complex technology. Potential sources of systematic errors include array surface chemistry, microarray printing, labeling methods, hybridization parameters, image analysis and RNA isolation (9-11). The process to correct for systematic error, generally termed normalization, is introduced to correct the differences across different arrays in probe labeling, probe concentration, hybridization efficiency and potentially other factors.

Normalizing multiple arrays to allow quantitative follow-up analyses presents one of the great challenges in microarray data analysis. Many normalization methods have been proposed in literature, the popular ones include global normalization or linear regression (LR) (12), Loess normalization (13), invariant ranking (IR) method (14), quantile normalization (15), and iterative nonlinear regression (INR) method (16,17). Regardless of their large technical differences, two basic steps in these methods involve: (1) selection of reference genes for normalization and (2) choice of a linear or nonlinear regression function for normalization (11).

For instances, Affymetrix's global normalization method uses all the genes for normalization with a linear regression function; Loess normalization method also uses all the genes for normalization but with a nonlinear regression function derived from M-A plots (18). In contrast, IR and INR methods use a subset of genes (i.e., rank invariant genes) for deriving a nonlinear regression function for normalization (14,16), while quantile normalization uses all the genes but the transformation function is derived in such a way that makes the distribution for each array in a set of arrays the same (15). In addition, housekeeping genes were used in the past for normalization under the assumption that they are constantly expressed genes (19), while in fact the expression levels of housekeeping genes can vary significantly (20). Exogenous control genes can also be used for normalization, and many reports have supported that it is an excellent and universally applicable normalization strategy (21).

Profound effect of normalization has been found on detection of differentially expressed genes and classification of phenotypes (22,23). Hoffmann *et al.* employed four different normalization methods and all possible combinations with three statistical algorithms for detection of differentially expressed genes (22). They found that the influence of normalization is significantly higher than that of three subsequent statistical analysis procedures examined. Hua *et al.* used a model-based approach to generate synthetic gene expression values for studying normalization procedures' impact on classification performance. Their experiment results demonstrated that normalization could have a significant benefit for classification under difficult experimental conditions (23).

In this paper, we report the experimental results from a comparison study of four normalization methods, namely, LR, Loess, IR and INR. We tested the normalization methods on three real and representative microarray data sets and evaluated their performance in terms of variance reduction and fold-change preservation. The performance in variance reduction mainly reflects the ability of a normalization method in correcting system error to make consistent gene expression measurements across multiple arrays. The performance of fold-change preservation, on the other end, shows the ability of a cross phenotype normalization method to reveal true phenotypic changes in gene expression measurements, which is critical to follow-up analyses to detect differentially expressed genes and classify different phenotypes. Note that we use the term "cross phenotype normalization" here to emphasize the importance of fold-change preservation. Between different phenotypes, the gene expressions are more diverse than within a phenotype. Hence, the normalization based on whole gene population (like LR method and Loess method) will introduce even large variance, impacting the fold-changes of the differentially expressed genes particularly. INR and IR are two methods that are based on invariantly expressed genes (IEGs) for normalization, although different ways are used to select the IEGs. We believe that for cross phenotype normalization, this is an important strategy to combat the large variance between different phenotypes.

3. NORMALIZATION METHODS

There have been many approaches proposed to normalize microarray gene expression data. Even though a variety of normalization strategies exit, we can categorize different normalization methods into three practical approaches: (1) global approach, (2) invariant gene approach and (3) exogenous control gene approach. The global approach is based on the assumption that the total mass of mRNA per cell is constant. Consequently, the total integrated intensity across all the genes should be roughly same in any two arrays. Rather than using all the genes, a subset of non-differentially expressed genes, i.e., invariantly expressed genes (IEGs), would be a good choice for normalizing microarray data across conditions. In contrast, using exogenous control genes for normalization is a universally applicable strategy since it does not depend on the assumptions like the ones described above. The use of exogenous control genes to normalize microarray data, while technically the most complex to set up and calibrate, may provide the best strategy for refining normalization methods.

In this section, we will focus on global and invariant gene approaches and give a detailed description of each method. Specifically, we will describe the algorithms of three global approaches - linear regression (LR), Loess regression and quantile normalization; and two invariant gene approaches - invariant ranking (IR) and iterative nonlinear regression (INR).

3.1. Linear regression

Linear regression (LR) method is the most commonly used normalization approach in large-scale gene expression analysis. This LR method is also referred to as "global scaling" in Affymetrix's analysis tools (12). The LR method can be described as follows for performing normalization in the probe level. A baseline array is first chosen; in practice, the array with median intensity is a reasonable choice for normalization. All other arrays are then normalized to this baseline array by some scaling factors estimated by linear regression. If $x_{k,baseline}$ are the intensities of the baseline array a $x_{k,i}$ is an array other than the baseline array (whe $k = 1,...,p$ represents the probe), the scaling factor β can be obtained via a common least-square minimization procedure, i.e., by minimizing the following mean squared error between x_i and its linear regression estimate $\hat{x}_i$

$$E\{(x_i - \hat{x}_i)^2\} = \frac{1}{p}\sum_{k=1}^{p}(x_{k,i} - \beta x_{k,baseline})^2 \quad (1)$$

As pointed out in (24), the linear regression method implicitly rests on the assumption that the amount of mRNA per cell is constant. However, this assumption is theoretically and practically questionable for several reasons like due to gene-dependent multiplicative errors and not the whole genome covered by an array (24).

3.2. Loess regression

Microarray expression levels may have large dynamic range that will cause scanner systematic deviations such as nonlinear response at lower intensity range and saturation at higher intensity. Although data falling into these ranges are commonly discarded for further analysis, the transition range, without proper handling, may still cause some significant error in differential expression gene detection. To account for this deviation, locally weighted linear regression (Loess) is regularly employed as a normalization method for such intensity-dependent effects (18).

This approach is based upon the idea of the *M* versus *A* plot (i.e., M-A plot), where *M* is the difference in log expression values and *A* the average of log expression values (25). For any two arrays (denote as *i*-th array and *j*-th array) with probe intensities $x_{k,i}$ and $x_{k,j}$ (where $k = 1,...,p$ represents the probe), respectively, we calculate *M* and *A* as following:

$$M_k = \log_2(\frac{x_{k,i}}{x_{k,j}}) \text{ and } A_k = \frac{1}{2}\log_2(x_{k,i}x_{k,j}) \quad (2)$$

A normalization function can be obtained by fitting this M-A plot using Loess regression. With the fitted normalization function $\hat{M}_k$, the normalization adjustment is $M_k^{'} = M_k - \hat{M}_k$ in order to be close to $M_k = 0$ axis. Thus, adjusted probe intensities for *i*-th and *j*-th arrays are given as

$$x_{k,i}^{'} = 2^{A_k + M_k^{'}/2} \text{ and } x_{k,j}^{'} = 2^{A_k - M_k^{'}/2} . \quad (3)$$

To deal with more than two arrays, the method can be further extended to look at all distinct pair-wise combinations. For each pair of arrays, we perform the Loess regression fitting in M-A plot and record the adjustment accordingly. After having performed on all distinct pairs, we have adjustments $M_{k,m,n}^{'}$ for each array *m* with respect to arrays, $n = 1,...,m-1,m+1,...,N$, where *N* is the total number of arrays. We then calculate the average adjustment to be applied to array *m*:

$$\bar{M}_{k,m}^{'} = \frac{1}{N-1}\sum_{n=1;n\neq m}^{N} M_{k,m,n}^{'} . \quad (4)$$

This process can be performed iteratively, which was properly termed as cyclic Loess regression method in (15). Although it seems to be a time consuming process, as reported in (15), the process will be stopped in a few iterations when the adjustments to be applied become small enough.

3.3. Invariant ranking

Contrast to the three normalization methods described in previous subsections, invariant ranking (IR) method differs in selecting a subset of genes for normalization (14). In particular, a subset of non-differentially expressed genes, i.e., invariantly expressed genes (IEGs), is identified by an iterative ranking method to estimate the normalization function. The method can be described as follows. For any two arrays, as in previous subsection, denoted as *i*-th array and *j*-th array with probe intensities $x_{k,i}$ and $x_{k,j}$ ($k = 1,...,p$ for the probe), we will first rank the probe in two arrays according to its intensity, denote the ranks as $r_{k,i}$ and $r_{k,j}$ for probe *k* in *i*-th and *j*-th arrays, respectively. We then calculate the rank difference between the two arrays, i.e., $d_k = r_{k,i} - r_{k,j}$ and normalize the rank difference as following:

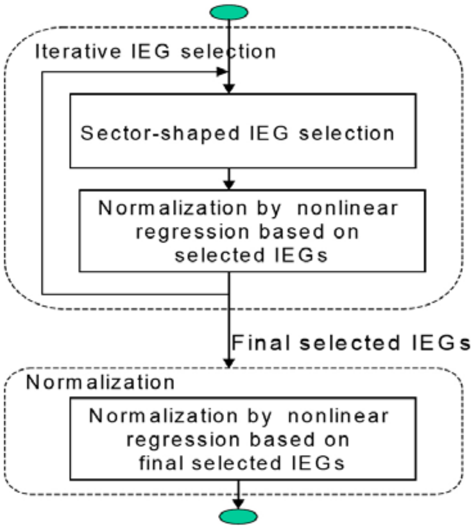

Iterative IEG selection
Sector-shaped IEG selection
Normalization by nonlinear regression based on selected IEGs
Final selected IEGs
Normalization
Normalization by nonlinear regression based on final selected IEGs

Figure 1. A block diagram of the normalization method by iterative nonlinear regression.

$$D_k = \frac{2|r_{k,i} - r_{k,j}|}{(r_{k,i} + r_{k,j})}. \tag{6}$$

The invariant set of genes is selected by comparing D_k to a threshold R_k given by

$$R_k = \frac{L(r_{k,i} + r_{k,j}) + H(2p - r_{k,i} - r_{k,j})}{2p}, \tag{7}$$

where, L and H are the rank difference thresholds for the low and high ends of the difference intensity range. As we can see, R_k is the linearly interpolated threshold between two ending thresholds (i.e., L and R). Specifically, probe k is included in the invariant set, if $D_k < R_k$; otherwise excluded from the invariant set. This selection process repeats, taking the current invariant set as input, until all normalized rank differences meet the threshold criteria. Once the final invariant set has been selected, the normalization function can be estimated by applying a nonlinear fitting technique, smoothing splines with generalized cross validation (GCVSS) (26), to the invariant set.

3.4. Iterative nonlinear regression

In this subsection, we describe INR normalization method in details. Figure 1 illustrates the block diagram of INR method consisting of two basic steps: (1) iterative IEG selection and (2) nonlinear regression normalization. As we can see, IEG selection is based on an iterative procedure that alternatively selects control genes (IEGs) and estimates nonlinear regression function for normalization. The final set of IEGs will be obtained when the iterative IEG selection procedure converges and subsequently, a nonlinear regression function will be estimated based on these IEGs. Next, we will describe the iterative IEG selection procedure and the outline of INR algorithm.

Different from most existing methods, INR normalization method relies on IEGs that can be selected iteratively by sector-shaped nonlinear regression (16). Specifically, we have developed an INR algorithm that alternatively selects IEGs and estimates normalization regression function. In an ideal case, i.e., without systematic errors, IEGs are the genes whose expression ratios are close to *1* between two microarray experiments, defined by the following equation mathematically:

$$\frac{1}{1+\delta} \le \frac{s_{\text{floating}}(i)}{s_{\text{reference}}(i)} \le 1+\delta, \tag{8}$$

where $s_{\text{reference}}$ and s_{floating} represent the expression levels of the reference (baseline) array and the floating array (i.e., the array to be normalized), respectively; δ is a pre-defined small threshold, and *i* is the gene index. Figure 2. shows an example of IEGs (as defined by Eq. (8)) in a scatter plot of two arrays, which reveals a sector-shaped distribution of IEGs.

Microarray data normalization aims to find a mapping function between the gene expression levels obtained from two samples or experiments. Mathematically, the gene expression levels in a floating array ($\hat{s}_{\text{floating}}$) can be modeled as a nonlinear regression function of the raw expression levels (s_{floating}) embedded with some systematic errors: $\hat{s}_{\text{floating}} = f(s_{\text{floating}})$ [14]. When the true IEGs are known or can be identified, we can estimate the nonlinear regression function by minimizing the mean squared error (MSE) between the expression levels in floating and reference arrays:

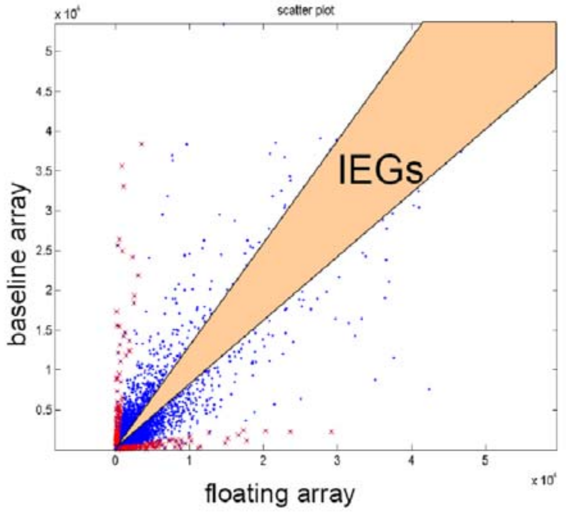


Figure 2. IEGs distributed within a sector-shaped region shown in scatter plot.

$$\varepsilon = \frac{1}{N_{IEG}} \sum_{i=1}^{N_{IEG}} \left[\frac{f^{-1}(\hat{s}_{\text{floating}}(i))}{s_{\text{reference}}(i)} - 1 \right]^2 , \quad (9)$$

where N_{IEG} is the number of IEGs and $s_{\text{reference}}(i)$ is the expression level of a particular IEG in a reference array. The popular forms of the nonlinear regression function include polynomials and smoothing splines. In particular, we have used the following three forms in the implementation - quadratic polynomials, cubic polynomials and smoothing splines with generalized cross validation (GCVSS) (26). It seems that cubic polynomials possess some advantage over quadratic polynomials and GCVSS, due to the accuracy in model fitting and low computational complexity in model parameter estimation.

In this paper, we describe an iterative procedure to find IEGs for nonlinear normalization as follows (15,16). The procedure repeats the following two steps until it converges: (1) selecting IEGs from a sector-shaped region in scatter plot of the floating and reference arrays; and (2) normalizing the floating array using the estimated nonlinear regression function based on selected IEGs (see Figure 1). Initially, we use a relatively large sector for selecting potential IEGs. For instance, we can start with using all the genes as IEGs (i.e., using a 90-degree sector angle), and perform an initial normalization accordingly. We then gradually decrease the angle of the sector-shaped region and select a new set of IEGs for normalization. The iterative procedure continues until there is no significant change in the content of IEGs and the estimated regression function converges to a 45-degree straight line (i.e., $f(s) = s$). Figure 3 illustrates the iterative process of IEG selection as the size of the sector decreases. The rationale of this approach lies in that after each normalization iteration the true IEGs shall move closer to a narrow sector around the 45-degree line as shown in Figure 2. Our numerical experiments have provided compelling evidence in support of such an iterative IEG selection scheme.

4. EXPERIMENTAL RESULTS

We have implemented several normalization methods including LR, Loess, IR and INR algorithms in C/C++ and integrated the modules into dChip software (27). In addition, all four above-mentioned methods have been implemented in a way that normalization can be carried out either at probe level for oligonucleotide array data or at gene level for cDNA array data. When carried out at probe level, we only use perfect match (PM) probes to select IEGs for normalization. Note that this is consistent with the implementation of iterative ranking (IR) method (14), but different from Bolstad's implementation where both PM and mismatch (MM) probes are used for invariant probe selection (15).

4.1. Data Sets and expression measurement

We used three data sets in our experimental tests - a dilution study from GeneLogic, a muscular dystrophy (MD) profiling study from Children's National Medical Center (CNMC), and a non-biological variability study from the Consortium for Functional Glycomics (CFG). The dilution data set was made available to the public specifically for comparison between different normalization methods (28). A total number of 60 arrays were acquired by Affymetrix's 75 HG-U95A microarrays to study the dilution/mixture effect of two sources of RNA from human liver tissue and central nervous system (CNS) cell line. The CNMC's MD data set with 125 arrays was acquired by Affymetrix's GeneChip (U133A) microarrays to study different types of muscular dystrophy (29). The Consortium for Functional Glycomics (CFG) has acquired 32 microarrays using custom-designed Glyco-gene Chips for assessment of sources of non-biological (technical) variability (30). Variables examined in the processing of RNA samples and gene chips include: (a) technician extracting RNA, (b) RNA isolation, (c) DNAse treatment of RNA, (d) biotin labeling of cRNA, and (e) day of hybridization to GLYCOv1. All samples in this study were C57/Bl mouse brain RNA. For all three data sets, the gene expression measurements were obtained using Affymetrix's Microarray Suite 5.0 (MAS 5.0) probe set interpretation algorithm (12), although other algorithms like robust multiple-array average (RMA) (31) and model-based expression index (MBEI) (32) can also be used.

4.2. Normalization plots

Figure 4 – Figure 7 show some typical results of the normalization methods when applied to GeneLogic's data set on dilution study. In the experiment, we chose an array (94407hgu95a11) as the baseline array since it is of median intensity among all arrays. The upper row in Figure 4 shows a second array (94394hgu95a11) normalized to the baseline array, while the lower row showing a third array (94420hgu95a11) before and after normalization. In a similar fashion, Figure 5 shows the normalization results using Loess regression method. As we can see, after normalization, the M-A plot of an array and baseline array is centered on $M = 0$. Figure 6 and Figure 7 show the results using IR and INR methods, respectively. Both IR and INR methods estimated nonlinear regression functions based on the selected IEGs as shown in Figure 6(b) and

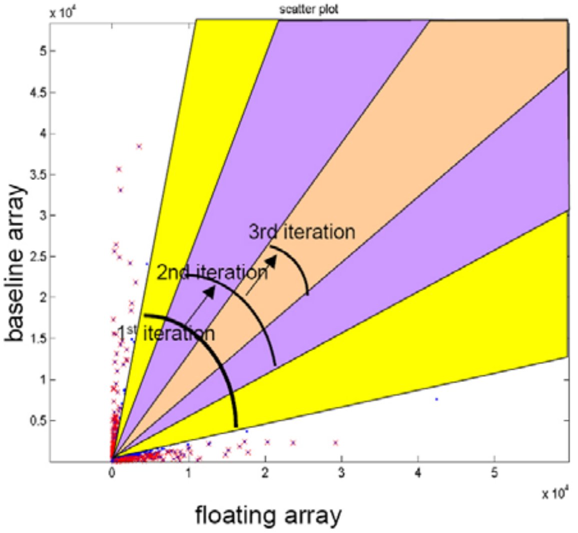


Figure 3. Iterative sector-shaped IEG selection by reducing the sector angle gradually.

Figure 7(b) (i.e., the red points). Evidently, as we can see from the figure, INR method effectively moved the IEGs to the 45-degree sector after normalization (Figure 7(c)).

Figure 8 shows examples of the iterative IEG selection process when applied to CNMC's MD data set. A large sector was initially used for IEG selection and regression function estimation. As iteration goes on, the sector was gradually narrowed down since IEGs were expected to move closer to the 45-degree line after each interim normalization step. The final set of IEGs was obtained when the following two conditions met: (1) the selected IEGs differ little from those selected in the previous step, and (2) the estimated regression function is close to the 45-degree line in scatter plot. Figure 9 shows the normalization result of INR, showing scatter plots of two MD arrays prior to normalization, final selected IEGs, and normalized MD arrays, respectively.

4.3. Performance comparison

To compare the performance of normalization methods such as LR, Loess, IR and INR methods, we used the following two criteria to quantitatively assess whether one method outperforms the other (14): (1) lower variance of expression level across replicated arrays, and (2) preservation of true fold-change in controlled realistic simulations. As discussed in (14), the first criterion ensures that genes known to have identical expression levels shall remain or incline to being identically expressed after normalization. The second criterion ensures that the first criterion is not achieved at the expense of destroying the very biological variations the technology aims to detect. Note that other criteria such as bias comparison based on spike-ins are also valuable to assess the performance of a normalization method under consideration (15).

4.3.1. Variance comparison

In GeneLogic's dilution study, there are 30 arrays for each RNA source (Liver or CNS) with 6 different masses of cRNA (1.25, 2.5, 5.0, 7.5, 10.0, and 20.0 µg). Each dilution level was hybridized on HG-U95A chips and then scanned by 5 different scanners as replicate measurements. This data set is ideal for performance comparison of different normalization methods, since non-biological variability (or systematic errors) was purposely introduced through replicates and dilutions, while the goal of normalization is to correct these system errors so that multiple arrays can be further analyzed for the problem being studied.

We used two sets of the 60 arrays of dilution study and one set of non-biological variability study from CFG for our variance comparison; the first set consisting of 30 arrays of liver, the second set consisting of 30 arrays of CNS, and the third set consisting of 32 arrays of C57/Bl mouse brain RNA. The following normalization methods were applied to the data sets: (1) LR method, (2) Loess method, (3) IR method and (4) INR method. After having normalized the arrays by these normalization methods respectively, we calculated expression measurements for each probe set on each array using MAS 5.0. We then computed the mean and variance of the expression measurements across all 30 arrays in each set. For variance comparison, we performed a pair-wise comparison between all four normalization methods. For any two methods (e.g., INR against IR), we counted the number of probe sets that have a larger variance of expression measurements using INR than that using IR. The percentage of the probe sets with larger variance was then calculated and used to assess the method's performance according to Criterion 1 (14).

Figure 10 shows the results using the liver data set from GeneLogic's dilution study. As we can see, all four normalization methods significantly reduced the expression variance when compared to the raw data (denoted as "UN" in Figure 7). All these normalization methods, in overall, produce more consistent expression measurements across these 30 arrays. In particular, IR and INR methods outperformed LR method in reducing the variance of expression measure (only about 30% and 28% of probe sets having larger variance than that using LR method, respectively). Furthermore, INR method showed 68% of probe sets having less variance than that from IR method, i.e., only 32% of probe sets having larger variance than that of IR method.

Figure 11 and Figure 12 shows the variance comparison results on the CNS data set and mouse brain data set, which again confirmed similar observations: (1) INR method exhibited a much better performance than LR and Loess methods in keeping the expression measurements consistent; (2) INR method further reduced the expression variance compared to IR method.

4.3.2. Fold-change comparison

In order to conduct fold-change comparison, we have constructed two sets of controlled realistically simulated microarray data based on GeneLogic's dilution data set. We chose ten replicates and dilution arrays to begin with - five of them were the replicate arrays at 5µg mass of cRNA from liver tissue and the other five were at

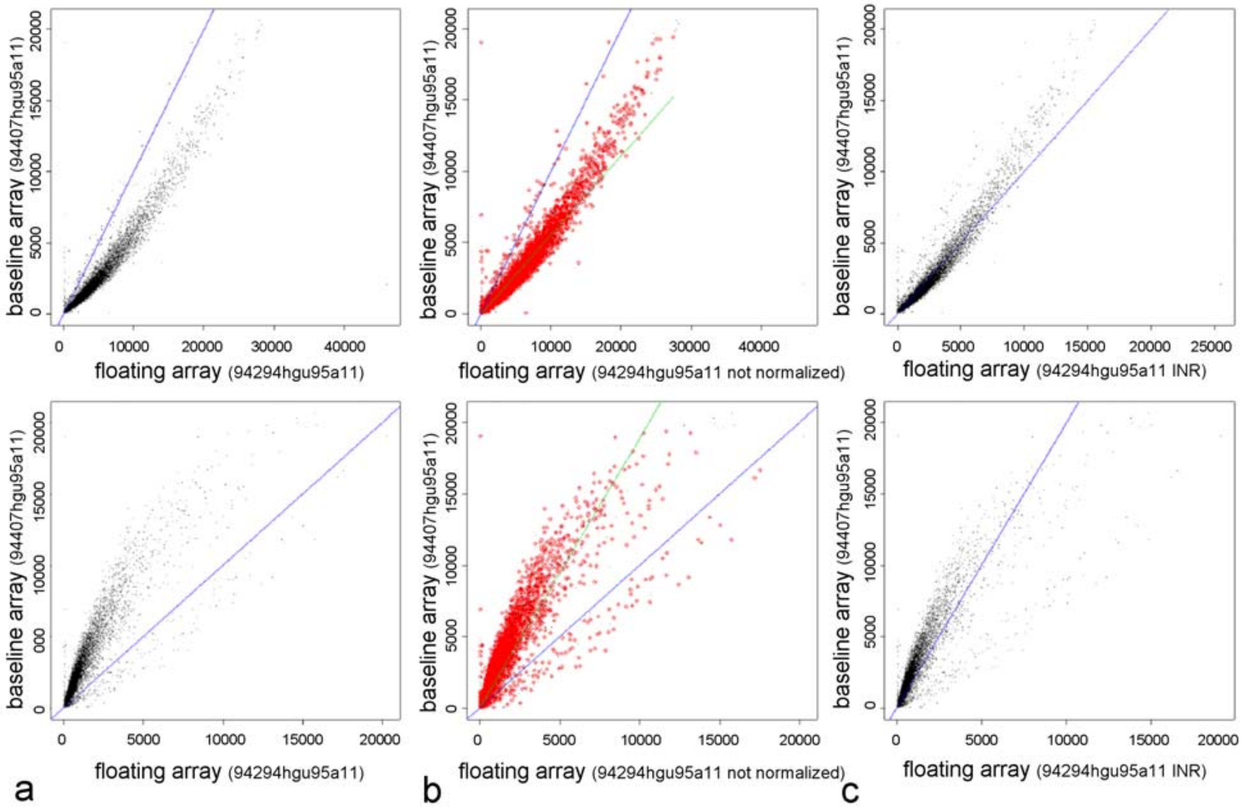


Figure 4. Scatter plots of the results using LR method. (Upper) sample array 94394hgu95a11 against the baseline array (94407hgu95a11); (Lower) sample array 94420hgu95a11 against the baseline array; (a) Before normalization, (b) estimated normalization function (green line), and (c) after normalization. The blue line is at 45-dgree. (Note: x-axis: floating array; y-axis: baseline array.)

10μg mass of liver cRNA. The simulated microarray data sets were constructed using the same procedure as originally designed by Schadt *et al.* (14). Below, we give a brief description of the procedure.

In the first set, 300 genes that were consistently detected as present across five low-intensity replicate arrays (5μg Liver cRNA) and 600 from high-intensity replicate arrays (10μg Liver cRNA) were randomly selected. Six sets containing 50 genes each for the low-intensity arrays and 100 genes each for the high-intensity arrays were then generated by a random selection process from the sets of 300 and 600 genes selected. The expression measurements of the selected genes in each of the six sets were then multiplied by 2.0, 0.5, 4.0, 0.25, 6.0, and 0.17, respectively, to simulate fold-changes between the samples. The ten original arrays (without modification) and ten modified arrays were used to compare the performance of normalization methods in preserving the controlled fold-changes. The same procedure was used to construct the second simulated data set consisting of ten replicates (5μg and 10μg of CNS cRNA) from dilution study of CNS. Similarly, the ten original arrays and ten modified arrays were used in the comparative experiments. Finally, a third data set was constructed in the same way using 10 mouse brain arrays from CFG as described in Section 3.1.

We tested the four different normalization methods (LR, Loess, IR and INR) on the same simulated data sets. After normalization, we calculated the fold-changes of the altered genes and computed the mean square errors (MSEs) between the observed and true fold-changes across replicates as follows:

$$\varepsilon_{\text{fold\_change}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{R}_i - R_i^0)^2, \tag{10}$$

where N is the number of arrays being modified (in this case, N = 10); R_i^0 is the true fold change (i.e., ground truth) and $\hat{R}_i$ is the observed fold change after normalization. Again, we performed a pair-wise comparison between all four normalization methods. For any two methods (e.g., INR against IR), we counted the number of genes having larger $\varepsilon_{\text{fold\_change}}$ when using

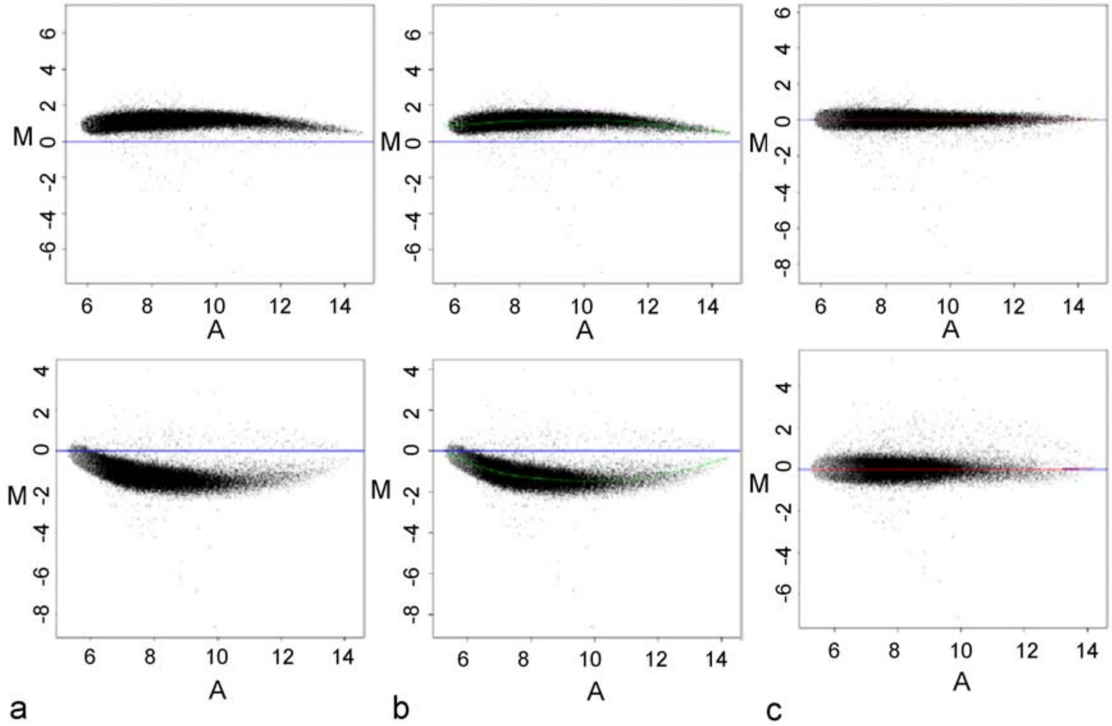


Figure 5. M-A plots of the results using Loess method. (Upper) sample 94394hgu95a11 against the baseline array (94407hgu95a11); (Lower) sample 94420hgu95a11 against the baseline array. (a) Before normalization, (b) estimated normalization function (green curve), and (c) after normalization. The blue line is at M = 0. (Note: x-axis: floating array; y-axis: baseline array.)

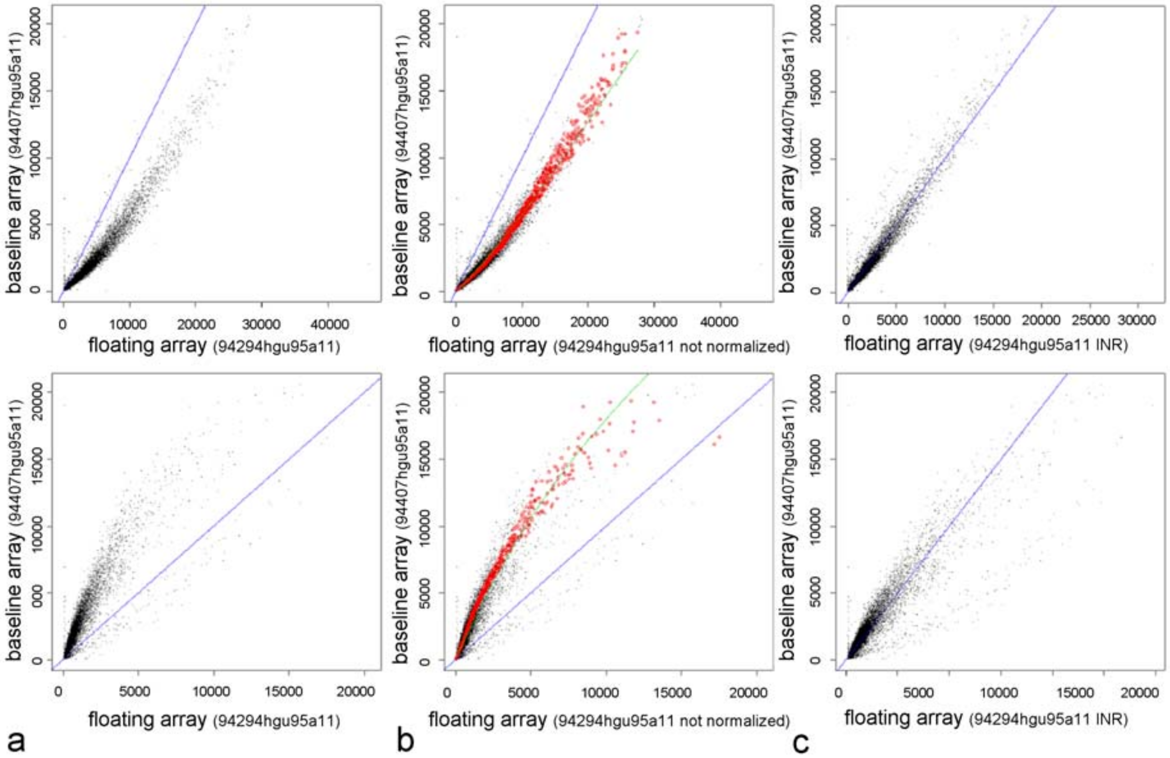


Figure 6. Scatter plots of the results using IR method. (Upper) sample 94394hgu95a11 against the baseline array (94407hgu95a11); (Lower) sample 94420hgu95a11 against the baseline array. (a) Before normalization, (b) estimated normalization function (green curve), and (c) after normalization. The red dots are the selected IEGs and the blue line is at 45-dgree. (Note: x-axis: floating array; y-axis: baseline array.)

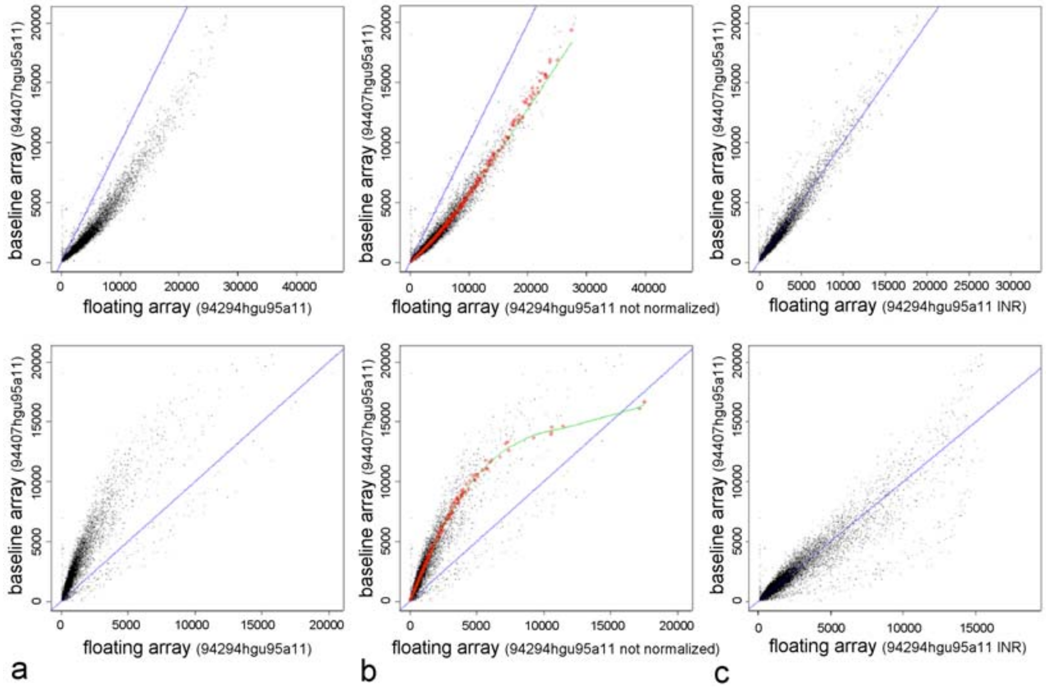


Figure 7. Scatter plots of the results using INR method. (Upper) sample 94394hgu95a11 against the baseline array (94407hgu95a11); (Lower) sample 94420hgu95a11 against the baseline array. (a) Before normalization, (b) estimated normalization function (green curve), and (c) after normalization. The red dots are the selected IEGs and the blue line is at 45-dgree. (Note: x-axis: floating array; y-axis: baseline array.)

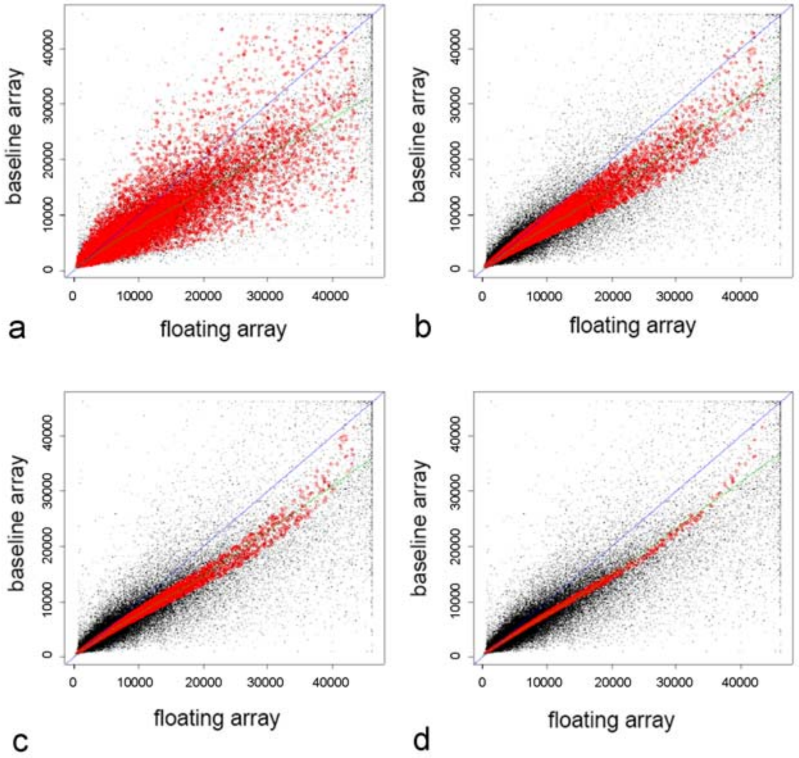


Figure 8. Iterative IEG selection by nonlinear regression - selected IEGs are in red: (a) initial IEGs (i.e., all the PM probes), (b) selected IEGs after 5 iterations, (c) selected IEGs after 10 iterations, and (d) final selected IEGs.

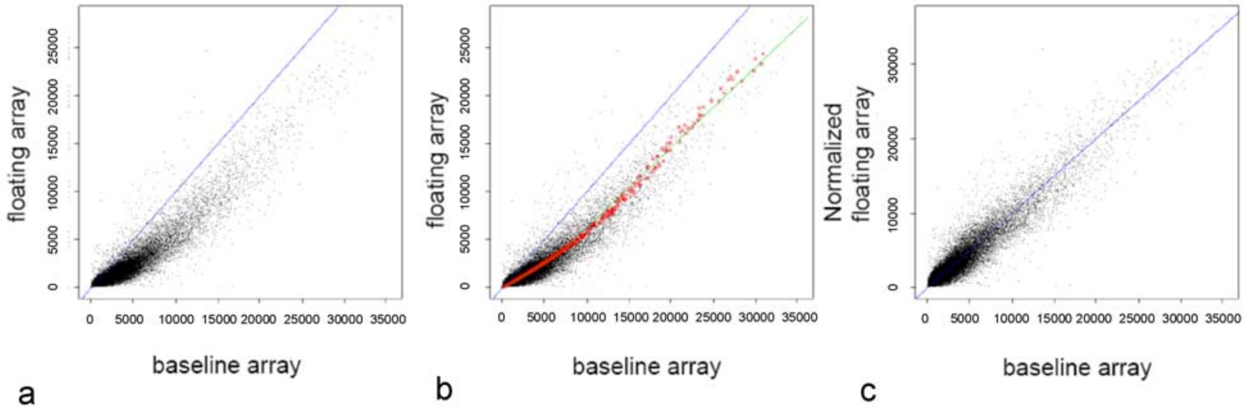
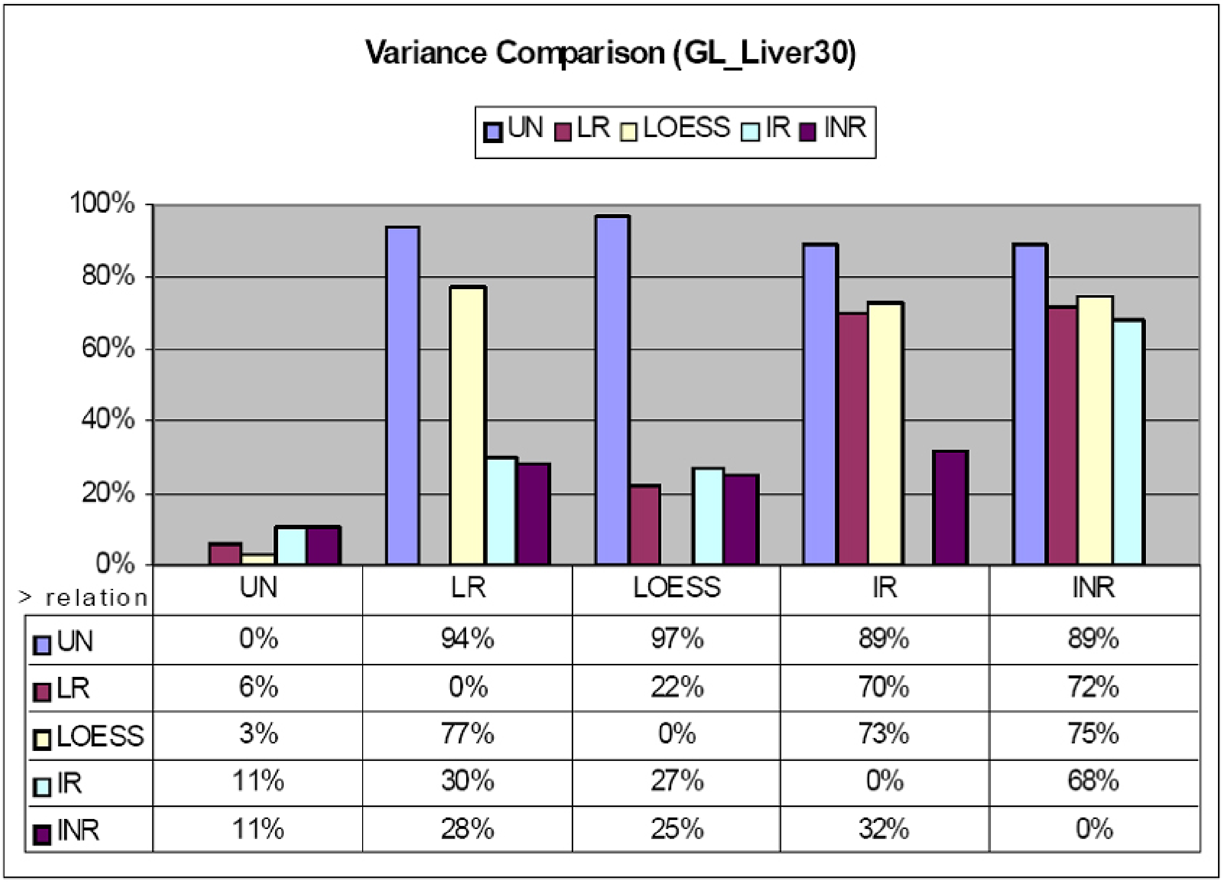
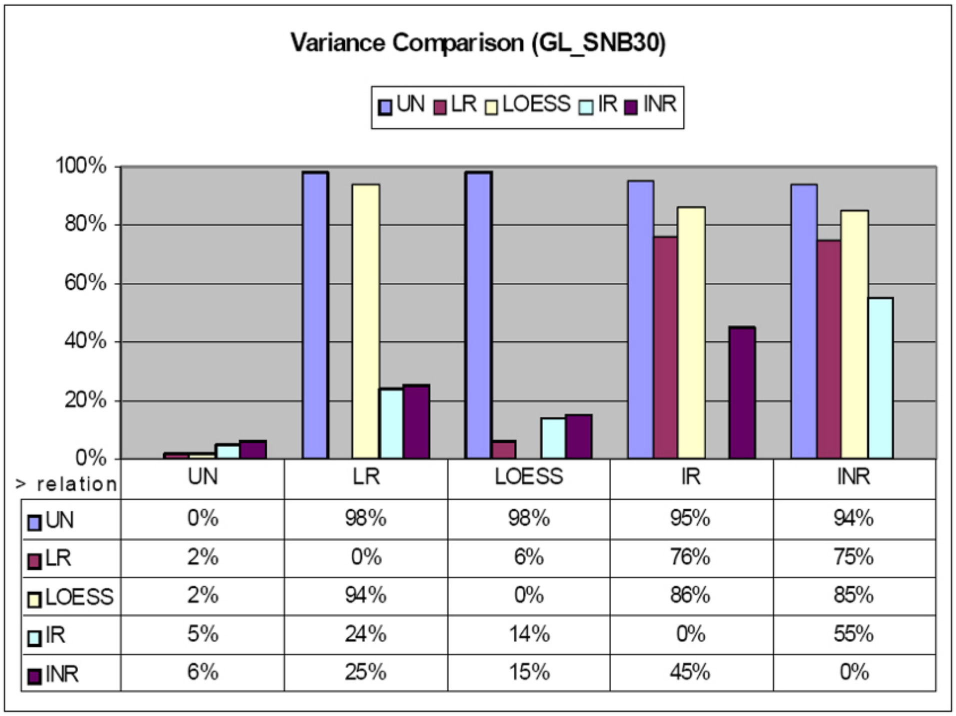


Figure 9. Normalization by INR method - an example of CNMC's MD data set: (a) scatter plot of unnormalized arrays, (b) selected IEGs for normalization, and (c) scatter plot of normalized arrays. The blue line indicates the 45-degree line.



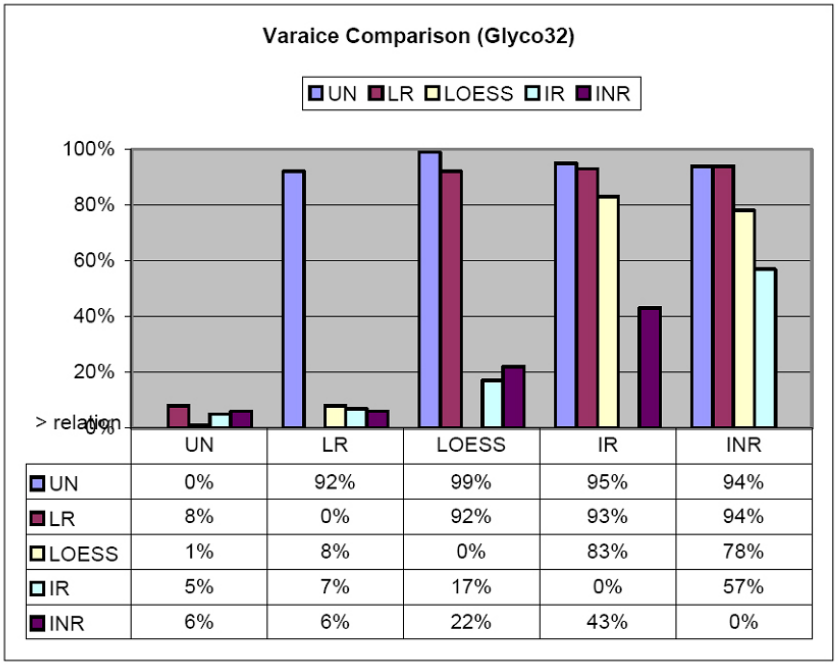
> relation	UN	LR	LOESS	IR	INR
■ UN	0%	94%	97%	89%	89%
■ LR	6%	0%	22%	70%	72%
□ LOESS	3%	77%	0%	73%	75%
□ IR	11%	30%	27%	0%	68%
■ INR	11%	28%	25%	32%	0%

Figure 10. Variance comparison using GeneLogic dilution data set (Liver). Four normalization methods, (1) LR, (2) Loess, (3) IR and (4) INR, are compared in terms of expression variance reduction. The normalization results are also compared with the unnormalized arrays (denoted as UN in the figure). The table should be interpreted as in the following example: (INR, LR) = 28% means that with INR method, only 28% of the genes are of larger expression variance than that with LR method.



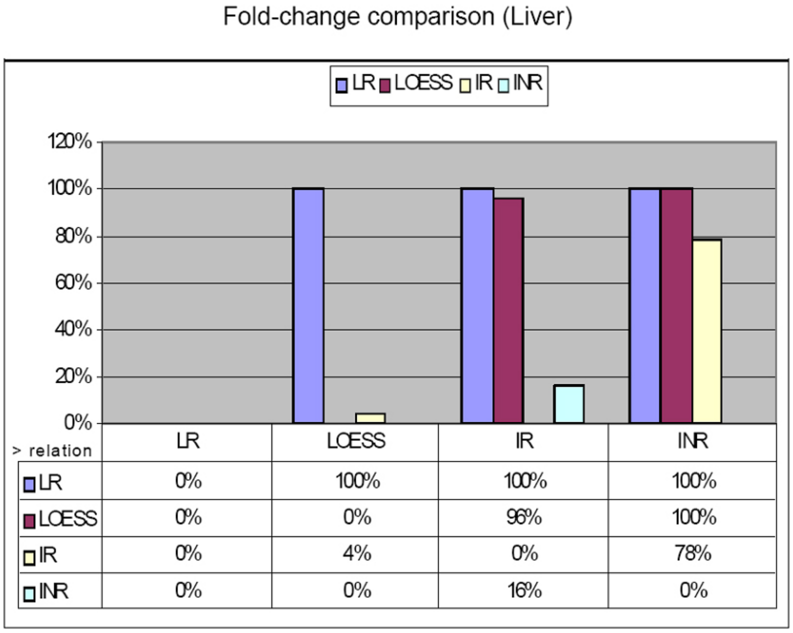
> relation	UN	LR	LOESS	IR	INR
UN	0%	98%	98%	95%	94%
LR	2%	0%	6%	76%	75%
LOESS	2%	94%	0%	86%	85%
IR	5%	24%	14%	0%	55%
INR	6%	25%	15%	45%	0%

Figure 11. Variance comparison using GeneLogic dilution data set (CNS). Four normalization methods, (1) LR, (2) Loess, (3) IR and (4) INR, are compared in terms of expression variance reduction. The normalization results are also compared with the unnormalized arrays (denoted as UN in the figure). The table should be interpreted as in the following example: (INR, LR) = 25% means that with INR method, only 25% of the genes are of larger expression variance than that with LR method.



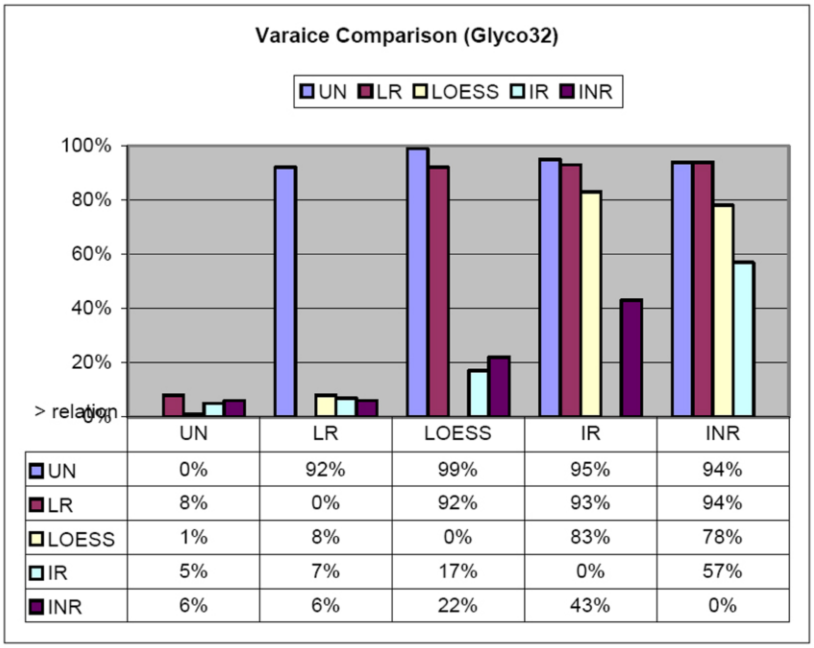
> relation	UN	LR	LOESS	IR	INR
UN	0%	92%	99%	95%	94%
LR	8%	0%	92%	93%	94%
LOESS	1%	8%	0%	83%	78%
IR	5%	7%	17%	0%	57%
INR	6%	6%	22%	43%	0%

Figure 12. Variance comparison using GFC mouse brain data set. Four normalization methods, (1) LR, (2) Loess, (3) IR and (4) INR, are compared in terms of expression variance reduction. The normalization results are also compared with the unnormalized arrays (denoted as UN in the figure). The table should be interpreted as in the following example: (INR, LR) = 6% means that with INR method, only 6% of the genes are of larger expression variance than that with LR method.



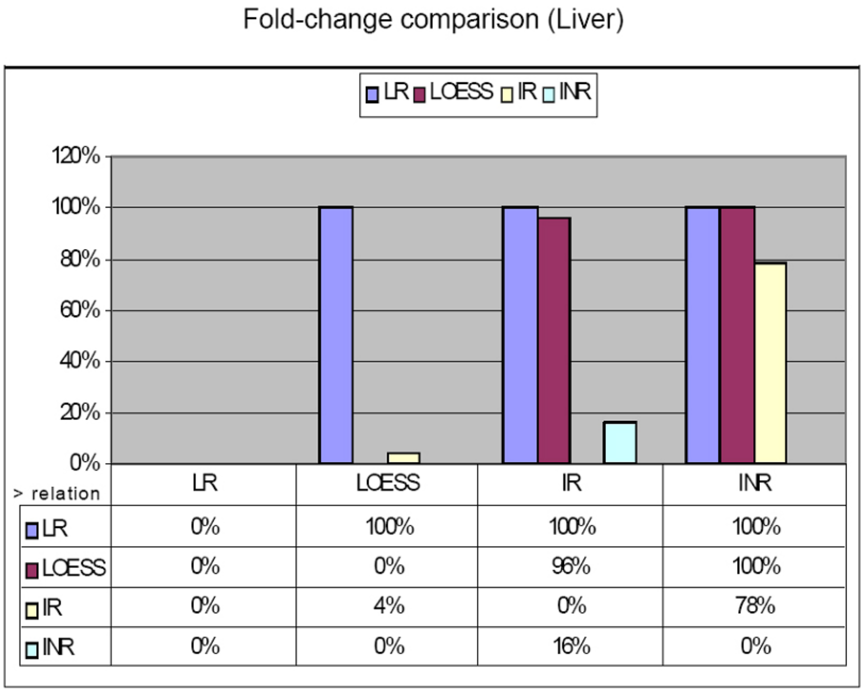
> relation	LR	LOESS	IR	INR
LR	0%	100%	100%	100%
LOESS	0%	0%	96%	100%
IR	0%	4%	0%	78%
INR	0%	0%	16%	0%

Figure 13. Fold-change comparison using GeneLogic dilution data set (Liver). Four normalization methods, (1) LR, (2) Loess, (3) IR and (4) INR, are compared in terms of fold change preservation. The table should be interpreted as in the following example: (INR, IR) = 16% means that with INR method, only 16% of the differentially expressed genes are of larger fold-change than that with IR method.



> relation	UN	LR	LOESS	IR	INR
UN	0%	92%	99%	95%	94%
LR	8%	0%	92%	93%	94%
LOESS	1%	8%	0%	83%	78%
IR	5%	7%	17%	0%	57%
INR	6%	6%	22%	43%	0%

Figure 14. Fold-change comparison using GeneLogic dilution data set (CNS). Four normalization methods, (1) LR, (2) Loess, (3) IR and (4) INR, are compared in terms of fold change preservation. The table should be interpreted as in the following example: (INR, IR) = 30% means that with INR method, only 30% of the differentially expressed genes are of larger fold-change than that with IR method.



> relation	LR	LOESS	IR	INR
LR	0%	100%	100%	100%
LOESS	0%	0%	96%	100%
IR	0%	4%	0%	78%
INR	0%	0%	16%	0%

Figure 15. Fold-change comparison using CFG's mouse brain data set. Four normalization methods, (1) LR, (2) Loess, (3) IR and (4) INR, are compared in terms of fold change preservation. The table should be interpreted as in the following example: (INR, IR) = 10% means that with INR method, only 10% of the differentially expressed genes are of larger fold-change than that with IR method.

INR than that using IR. The percentage of the genes with larger $\varepsilon_{\text{fold\_change}}$ was then calculated for assessing the performance according to Criterion 2 (14).

Figure 13 shows the comparison results of fold-change preservation on the first testing data set (Liver). The performance can be observed as follows. First, LR method was the worst one among all four normalization methods in preserving the authentic fold-changes. Second, Loess method was the second worst method in that it exhibited 96% of genes having larger $\varepsilon_{\text{fold\_change}}$ than that using IR method, and 100% of genes having larger $\varepsilon_{\text{fold\_change}}$ than that using INR method. Third, INR method gave the best performance in terms of fold-change preservation, only 16% of genes having larger $\varepsilon_{\text{fold\_change}}$ than that using IR method.

Figure 14 and Figure 15 show the comparison results on the second and third testing data sets (CNS and mouse brain, respectively). Among all four normalization methods, LR method was again the worst one in terms of fold-change preservation. As expected, INR method continued to show the best performance in preserving fold-changes, specifically, only 30% (or 25%) of the genes having larger $\varepsilon_{\text{fold\_change}}$ than that using IR method (or Loess method) in the CNS data set; 10% against IR method (or 0% against Loess method) in the mouse brain data set.

5. CONCLUSION

In this paper, we have reported a comparison study of four normalization methods, namely LR, Loess, IR and INR methods, for normalizing gene expression data. We tested the normalization methods on three real microarray data sets – GeneLogic's array data set for dilution study, CNMC's microarray data set for muscular dystrophy study and CFG's data set for non-biological variability study – to evaluate their performance. The experimental results have demonstrated that an improved performance can be obtained using normalization methods for correcting systematic errors. It becomes evident to us that correct selection of IEGs is the key to assure the success of any normalization method. Not like other methods (e.g., LR method, Loess method and quantile method), INR and IR methods are the only ones that perform the normalization based on IEGs selected via carefully designed procedures.

Specifically, we compared the performance of the above-mentioned normalization methods based on the following two criteria: (1) expression variance reduction and (2) fold-change preservation. From the experimental results, we have come to the conclusions that (1) LR method was the worst one among the four normalization methods tested on the data sets used in the experiments; and (2) INR method outperformed all other three methods (LR, Loess and IR methods) in reducing expression variance across replicates and preserving the fold-changes of targeted differentially expressed genes.

It is worth noting that there have been several other important comparison studies on microarray data normalization (14,15,32). Schadt *et al.* have compared the IR method with LR and GCVSS methods (14). They have demonstrated that the selection of invariantly expressed genes (IEGs) is very important for reducing the array variance without compromising the fold-change preservation. Since they used the same criteria (variance reduction and fold-change preservation) as those in this paper to evaluate the performance, the results from their study were consistent with our results. Furthermore, our method (i.e., INR) improves the performance with a novel IEG identification method (by iteratively normalizing the floating array so that the IEGs are moved to the 45-degree line in the scatter plot). Bolstad *et al.* also conducted a comparison study to evaluate several normalization methods like Loess, LR, IR and quantile methods (15). In their study, they used variance and bias as the criteria to evaluate the performance of each method; they also used GeneLogic's dilution data set for variance comparison, but used GeneLogic's Spike-in data set for bias comparison. They found that the quantile method gave a slightly better performance than the other methods; the LR method was the worst among the methods, which is consistent with the comparison result in this paper; however, the nonlinear method, IR, also did poorly for the spike-in regression in their study. Recently, Fujita *et al.* have also compared different methods, including Loess, splines, wavelets, kernel smoothing and support vector regression (SVR) to evaluate their performances using simulated microarray data (32). They have shown that the SVR method was favored for microarray normalization due to its robustness in estimating the normalization curve. The SVR method is essentially a nonlinear method using all the genes, but optimized to limit the fitting error while finding a linear mapping function as flat as possible (32).

In the future, we will further include those new approaches like SVR and quantile normalization to complete the performance evaluation of different normalization methods for microarray data normalization. In addition, we will use the simulated microarray data as generated in (32) and bench microarray data together to further assess the robustness of each method in terms of variance reduction and fold-change preservation. Finally, we also plan to further improve the INR method with the idea of SVR method to optimize the nonlinear regression function for better identifying the ISGs.

6. ACKNOWLEDGMENTS

This work was supported in part by the NIH under Grants CA109872, EB000830 and NS29525-13A; DOD/CDMRP under Grant BC030280. We thank C. Li and W. H. Wong at Harvard University for making dChip software, especially the implementation of their iterative ranking (IR) normalization method, available to us for comparison. We also thank the anonymous reviewers for their invaluable inputs that lead to several important improvements.

7. REFERENCES

1. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R., Downing, M. A. Caligiuri, C. D. Bloomfield and E.S. Lander: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537 (1999)

2. J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Lananyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 7(6), 673-679 (2001)

3. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* 34(2), 166-176 (2003)

4. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863-14868 (1998).

5. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, E. Dmitrovsky, E. S. Lander and T. R. Golub: Interpreting gene expression with self-organizing maps: Methods and application to hematopoeitic differentiation. *Proc. Natl. Acad. Sci. USA* 96, 2907-2912 (1999)

6. Y. Wang, L. Luo, M. T. Freedman and S. Y. Kung: Probabilistic principal component subspaces: A hierarchical finite mixture model for data visualization. *IEEE Trans. Neural Nets.* 11, 625-636 (2000).

7. S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander and T. R. Golub: Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* 98(26), 15149-15154 (2001).

8. J. Xuan, Y. Dong, J. I. Khan, E. Hoffman, R. Clarke and Y. Wang: Robust feature selection by weighted Fisher criterion for multiclass prediction in gene expression profiling. *Proc. Int'l Conf. Pattern Recognition* (2), 291-294 (2004)

9. A. Hartemink, D. Gifford, T. Jaakkola and R. Young: Maximum likelihood estimation of optimal scaling factors for expression array normalization. *SPIE BIOS* (2001)

10. G. C. Tseng, M. K. Oh, L. Rohlin, J. C. Liao and W. H. Wong: Issues in cDNA microarray analysis: quality filering, channel normalization, models of variations and assessment of gene effects. *Nucl. Acids Res.* 29, 2549-2557 (2001)

11. M. Bilban, L. K. Buehler, S. Head, G. Desoye and V. Quaranta: Normalizing DNA microarray data. *Curr. Issues Mol. Biol.* 4, 57-64 (2002)

12. Affymetrix : Affymetrix Technical Note Statistical algorithms description document. *Affymetrix, Inc.* (http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf) (2002)

13. Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai and T.P. Speed: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slides systematic variation. *Nucleic Acids Res.* 30(4), e15 (2002)

14. E. Schadt, C. Li., B. Eliss and W. H. Wong: Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* 84(S37), 120-125 (2001)

15. B. M. Bolstad, R. A. Irizarry, M. Astrand and T.P. Speed: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2), 185-193 (2003)

16. Y. Wang, J. Lu, R. Lee, Z. Gu and R. Clarke: Iterative normalization of cDNA microarray data. *IEEE Trans. on Info. Tech. in Biomedicine* 6(1), 29-37 (2002)

17. J. Xuan, E. Hoffman, R. Clarke and Y. Wang: Normalization of microarray data by iterative nonlinear regression. *Proc. the Fifth IEEE Symposium on Bioinformatics and Bioengineering*, 267-270 (2005)

18. J. Quackenbush: Microarray data normalization and transform. *Nature Genetics Suppl.* 32, 496-501 (2002)

19. E. Camerer, E. Gjernes, M. Wiiger, S. Pringle and H. Prydz: Binding of factor VIIa to tissue factor on karatinocytes induces gene expression. *J. Biol. Chem.* 275, 6580-6585 (2000)

20. J. B. Welsh, L. M. Sapinoso, A. I. Su, S. G. Kern, J. Wang-Rodriguez, C. A. Moskaluk, H. F. Frierson Jr and G. M. Hampton: Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.* 61, 5974-5978 (2001)

21. A. A. Hill, E. L. Brown, M. Z. Whitley, G. Tucker-Kellogg, C. P. Hunter and D. K. Slonim: Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol.* 2(12), 1-12 (2001)

22. R. Hoffmann, T. Seidl and M. Dugas: Profound effect of normalization on detection of differentially expressed genes in olgonecleotide microarray data analysis. *Genome Biology* 3(7), 1-11 (2002)

23. J. Hua, Y. Balagurunathan, Y. Chen, J. Lowey, M. L. Bittner, Z. Xiong, E. Suh and E. R. Dougherty: Normalization Benefits Microarray-Based Classification. *EURASIP Journal on Bioinformatics and Systems Biology* 2006, Article ID 43056, 13 pages, (2006)

24. A. Zien, T. Aigner, R. Zimmer and T. Lengauer: Centralization: A new method for the normalization of gene expression data. *Bioinformatics* 1(1), 1-9 (2001)

25. S. Dudoit, Y. H. Yang, M. J. Callow and T. P. Speed; Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Stat. Sin. 12(1), 111-139 (2002)

26. G. Wahba: *Spline methods for observational data,* CBMS-NSF regional conf. series in applied math. Philadelphia: SIAM (1990)

27. C. Li and W. H. Wong: DNA-Chip Analyzer (dChip). *The analysis of gene expression data: methods and software*. Edited by Parmigiani, G., Garrett, E.S., Irizarry, R. and Zeger, S.L. Springer (2003)

28. GeneLogic: Dilution/mixture datasets. http://www.genelogic.com (2002)

29. M. Bakay, Z. Wang, G. Melcon, L. Schiltz, J. Xuan, P. Zhao, V. Sartorelli, J. Seo, E. Pegoraro, C. Angelini, B. Shneiderman, D. Escolar, Y.-W. Chen, S. Winokur, L. Pachman, C. Fan, R. Mandler, Y. Nevo, E. Gordon, Y. Zhu, Y. Dong, Y. Wang and E.P. Hoffman: Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb–MyoD pathways in muscle regeneration. *Brain* 129, 996-1013 (2006)

30. The Consortium for Functional Glycomics (CFG): Glyco-gene Chip and microarrat data. URL: http://www.functionalglycomics.org/glycomics/publicdata/microarray.jsp

31. R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–64 (2003)

32. A. Fujita, J. R. Sato, O. Rodrigues Lde, C. E. Ferreira and M. C. Sogayar: Evaluating different methods of microarray data normalization. *BMC Bioinformatics* 7, 469-479 (2003)

Abbreviations: LR: linear regression, IR: invariant ranking, INR: iterative nonlinear regression, GCVSS: smoothing splines with generalized cross validation, IEG: invariantly expressed genes, MD: muscular dystrophy, MSE: mean square error

Key Words: Normalization; nonlinear regression; gene expression profiling; microarray data analysis; computational bioinformatics

Send correspondence to: Jianhua Xuan, Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, 4300 Wilson Blvd.,

Arlington, VA 22203, USA, Tel: 703-387-6057, Fax: 703-
528- 5543, E-mail: xuan@vt.edu

http://www.bioscience.org/current/vol2E.htm