An Iterative Nonlinear Regression Method for Microarray Data Normalization

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

[1]*Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA, USA*

[2]*Department of Oncology, Georgetown University School of Medicine, Washington, DC, USA*

[3]*Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC, USA*

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 novel normalization approach, namely iterative nonlinear regression (INR) method, which exploits concurrent identification of invariantly expressed genes (IEGs) and implementation of nonlinear regression normalization. The INR scheme features an iterative process that performs the following two steps alternatively: (1) selection of IEGs and (2) estimation of nonlinear regression function for normalization. We demonstrate the principle and performance of the INR approach on two real microarray data sets. As compared to major peer methods (e.g., linear regression method, Loess method and iterative ranking method), INR method shows an improved performance in achieving low expression variance across replicates and excellent fold-change preservation for differently expressed genes.

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, 2]. 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], rank invariant method [14], and quantile normalization [15]. 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 [9].

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 [16]. In contrast, rank invariant method uses a subset of genes (i.e., rank invariant genes) for deriving a nonlinear regression function for normalization, 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 [17], while in fact the expression levels of housekeeping genes can vary significantly [18]. Exogenous control genes can also be used for normalization, and many reports have supported that it is an excellent and universally applicable normalization strategy [19].

In this paper, we report a novel approach for microarray data normalization using an iterative nonlinear regression

*Address correspondence to this author at the Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA, USA;
E-mail: xuan@vt.edu

1874-1363/07

(INR) strategy. The basic idea of INR method is to couple the two key steps in the normalization procedure, i.e., (1) selection of invariantly expressed genes (IEGs) and (2) deriving a regression function for normalization, into an iterative search that effectively updates the selection of IEGs for normalization. We tested the INR method on two real and representative microarray data sets and evaluated INR performance in terms of variance reduction and fold-change preservation.

METHOD

In this section, we describe INR normalization method in details. Fig. (**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.

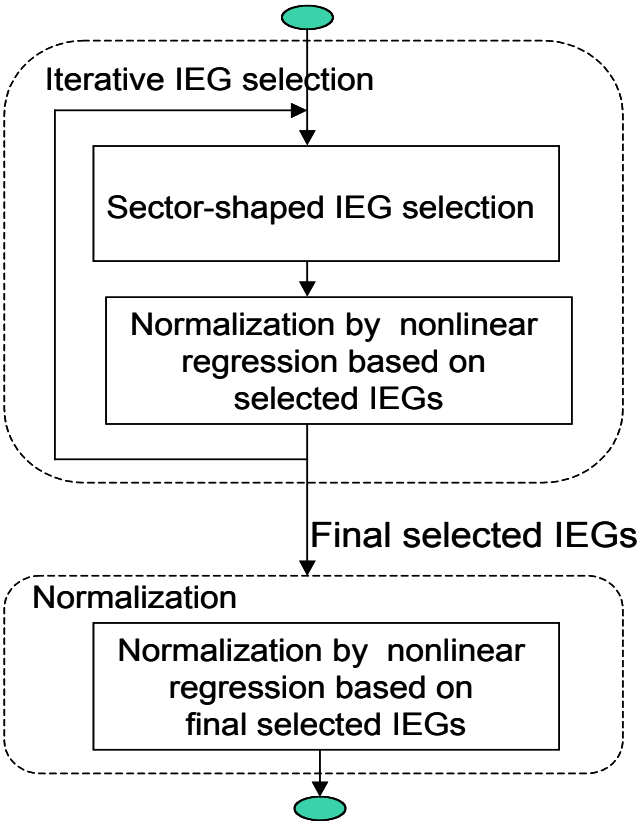


Fig. (1). Block diagram of the normalization method by iterative nonlinear regression.

ITERATIVE IEG SELECTION

Different from most existing methods, INR normalization method relies on IEGs that can be selected iteratively by sector-shaped nonlinear regression [20]. 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:

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

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. Fig. (**2**) shows an example of IEGs (as defined by Eq. (1)) in a scatter plot of two arrays, which reveals a sector-shaped distribution of IEGs.

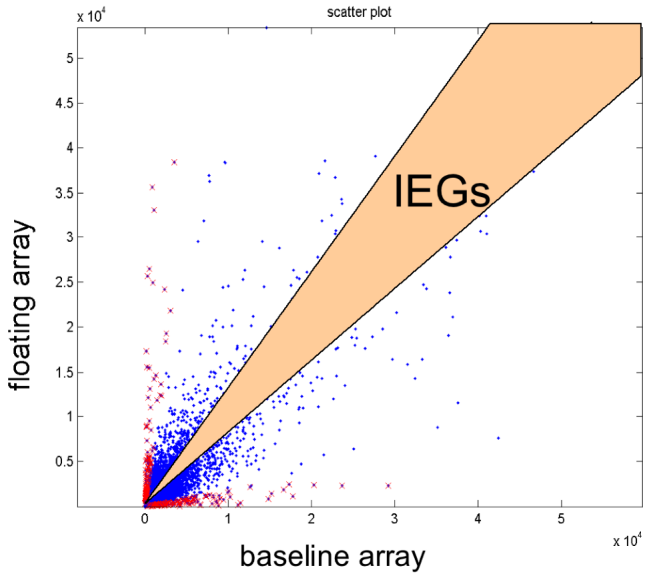


Fig. (2). IEGs distributed within a sector-shaped region shown in scatter plot.

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 (s_{floating}) can be modeled as a nonlinear regression function of the raw expression levels (s_{floating}) embedded with some systematic errors: $s_{\text{floating}} = f(s_{\text{floating}})$ [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:

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

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) [21]. 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.

Now we turn to the key question on how to find the true IEGs. In this paper, we devise an iterative procedure to find IEGs for nonlinear normalization as follows [20]. 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 Fig. (**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., $f(s) = s$). Fig. (**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 Fig. (**2**). Our numerical experiments have provided compelling evidence in support of such an iterative IEG selection scheme.

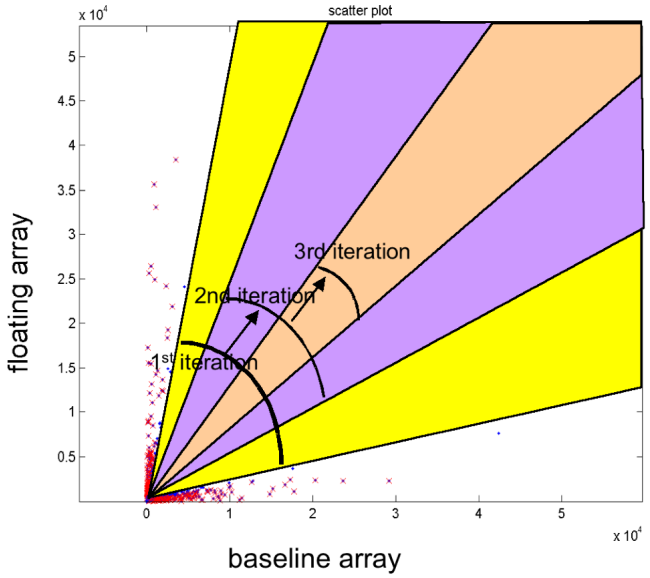


Fig. (3). Iterative sector-shaped IEG selection by reducing the sector angle gradually.

INR Algorithm

In this section, we will outline the INR algorithm that consists of three basic steps: (1) initialization, (2) iterative IEG selection, and (3) normalization by nonlinear regression. The outline of the algorithm is given below:

INR Algorithm Outline

1. Initialization:
   a. Perform an initial nonlinear normalization using all the genes (i.e., with a 90-degree sector angle).
2. Iterative IEG selection:
   a. Select IEGs within the defined sector-shaped region in the scatter plot;
   b. Estimate the nonlinear regression function (e.g., a cubic polynomial function) using the selected IEGs (by minimizing the MSEs of Eq. (2));
   c. Normalize the floating array using estimated nonlinear regression function;
   d. Decrease the size of sector-shaped IEG selection region (e.g., using a sector angle that is 90% of the previous sector angle);
   e. Go to 2-a, if either the selected IEGs different from previously selected IEGs or the regression function not approaching to a 45-degree line in scatter plot.
3. Normalization by nonlinear regression:
   a. Estimate the nonlinear regression function using the selected IEGs;
   b. Normalize the floating array by the estimated nonlinear regression function.

As for any nonlinear regression problem, it is generally not guaranteed to result in a unique solution; the algorithm may not converge to the solution or being stuck in a local minimum. However, the problem is alleviated in this special application as reasoned and justified below. (1) The degree of nonlinearity in microarray data normalization is relatively moderate; in our experiments, cubic polynomials are good enough to model the nonlinearity introduced by the systematic errors. (2) The proposed algorithm is an iterative approach that performs the following two steps alternatively: (a) selection of invariantly expressed genes (IEGs) and (b) estimation of nonlinear regression function. As the iteration goes on, the IEGs are moved closer and closer to a narrow sector around the 45-degree line (see Fig. (**3**)). Therefore, the fitting function prone to become a linear function gradually (at last, it becomes a 45-degree line ideally). In our implementation, we also check the change in number of IEGs to make sure that the algorithm will stop. We have tested the algorithm on a large number of microarrays for normalization. From our experience, it seems that (1) the algorithm does give us reasonable good solutions to the problems (even though the solutions may not be optimal); (2) the algorithm does converge to a solution after certain iterations.

RESULTS

We have implemented the INR algorithm in C/C++ and integrated the INR module into dChip software [22]. In addition, INR method has 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].

DATA SETS AND EXPRESSION MEASURMENT

We used two data sets in our experimental tests - the dilution experiment from GeneLogic and the muscular dystrophy (MD) profiling experiment from Children's National

Medical Center (CNMC). The dilution data set was made available to the public specifically for comparison between different normalization methods [23]. 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 [24]. For both data sets, the gene expression measurements were obtained using Affymetrix's MAS 5.0 probe set interpretation algorithm [12].

INR NORMALIZATION

Fig. (**4**) shows an example 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 that selected in the previous step, and (2) the estimated regression function is close to the 45-degree line in scatter plot. Fig. (**5**) shows the normalization result of INR, showing

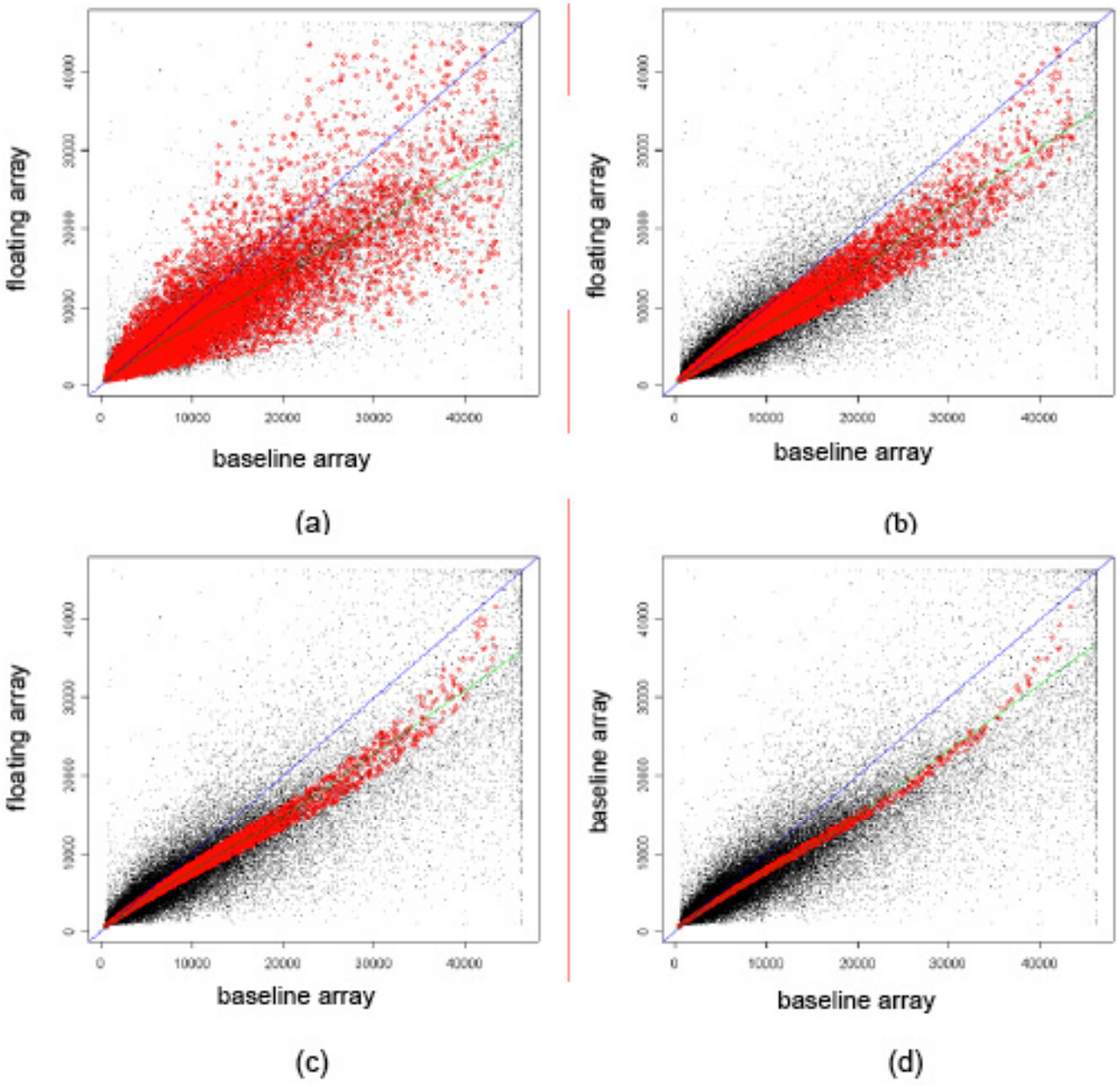


Fig. (4). Iterative IEG selection (the red dots are the selected IEGs and the black dots are non-IEGs). (**a**) Initial IEGs, (**b**) selected IEG after 5 iterations, (**c**) selected IEGs after 10 iterations, and (**d**) final selected IEGs. The green curve is the estimated nonlinear regression function using the selected IEGs. The blue line indicates the 45-degree line.

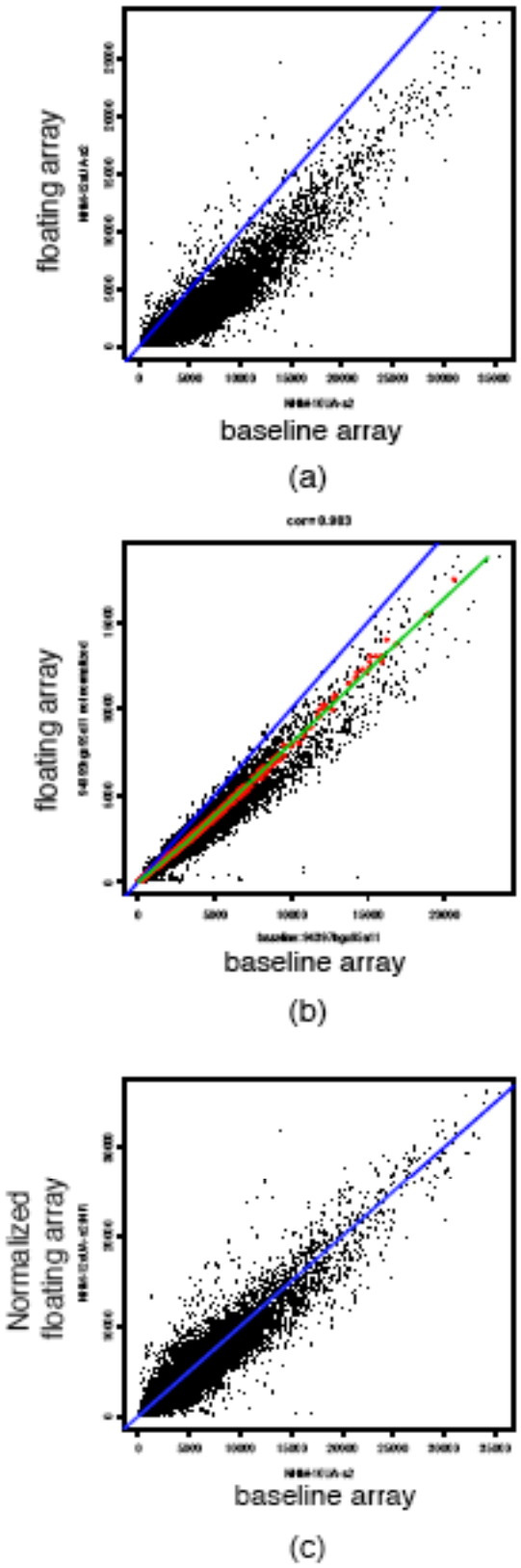


Fig. (5). Normalization by INR method - an example of CNMC's MD data set (the red dots are the selected IEGs and the black dots are non-IEGs). (**a**) scatter plot of unnormalized arrays (floating array *vs* baseline array), (**b**) selected IEGs for normalization (the green curve is the estimated nonlinear regression function using the selected IEGs), and (**c**) scatter plot of normalized arrays. The blue line indicates the 45-degree line.

scatter plots of two MD arrays prior to normalization, final selected IEGs, and normalized MD arrays, respectively.

Fig. (**6**) shows some typical results of INR as applied to normalizing GeneLogic's data set on dilution study. In the experiment, we chose an array (94397hgu95a11) as the baseline array since it is of median intensity among all arrays. Fig. (**6**) shows a second array (94394hgu95a11) normalized to the baseline array. The INR method estimated nonlinear regression functions based on the selected IEGs as shown in Fig. (**6b**) (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 (Fig. **6c**).

PERFORMANCE COMPARISON

To compare the performance of INR method with that of peer methods such as LR, Loess and IR 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].

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 for our variance comparison, the first set consisting of 30 arrays of liver and the second set consisting of 30 arrays of CNS. 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].

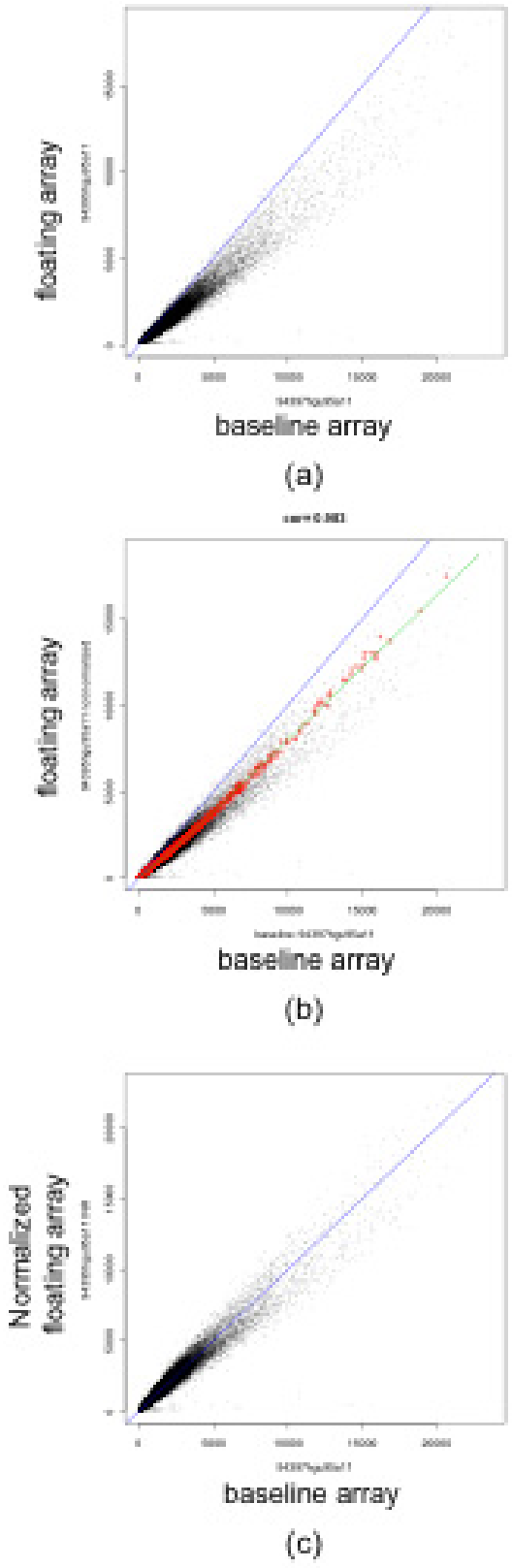
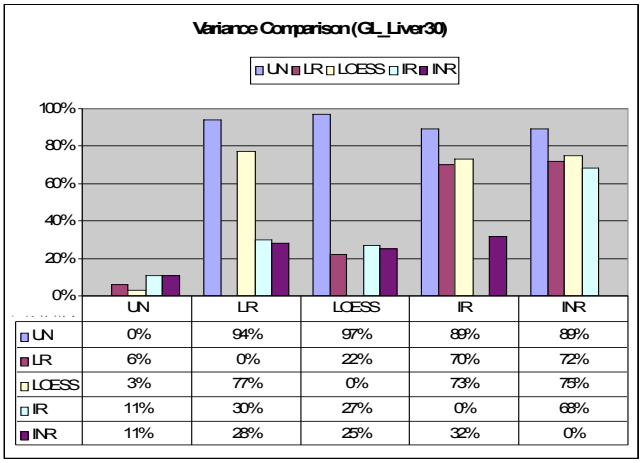


Fig. (6). Normalization by INR method - an example of GenLogic's dilution data set (the red dots are the selected IEGs and the black dots are non-IEGs). (**a**) scatter plot of unnormalized arrays (floating array *vs* baseline array), (**b**) selected IEGs for normalization (the green curve is the estimated nonlinear regression function using the selected IEGs), and (**c**) scatter plot of normalized arrays. The blue line indicates the 45-degree line.



	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%

Fig. (7). Variance comparison using GeneLogic dilution data set (Liver). Four normalization methods, (1) LR (Linear Regression), (2) Loess (Loess Regression), (3) IR (Iterative Ranking) and (4) INR (Iterative Nonlinear Regression), 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.

Fig. (**7**) 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 Fig. (**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.

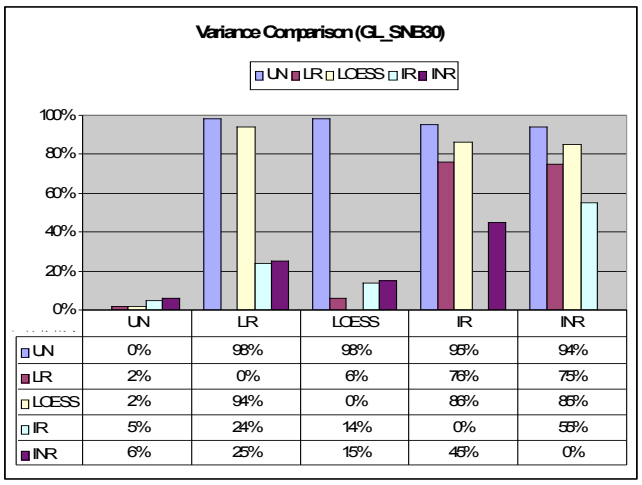
Fig. (**8**) shows the variance comparison results on the CNS data set, which again confirmed similar observations: (1) INR method exhibited a much better performance than LR and Loess methods in keeping the expression measurements consistent; (2) INR method further reduced the expression variance compared to IR method.

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 10μg mass of liver cRNA. The simulated microarray data sets were constructed using the same procedure as originally designed by Schadt *et al.* 2001[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.



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

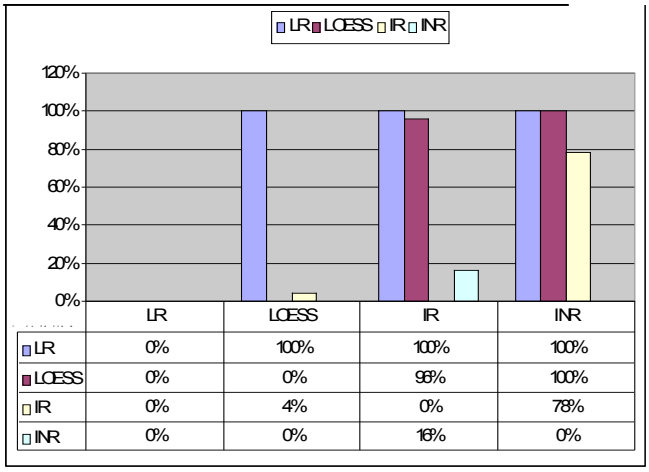
Fig. (8). Variance comparison using GeneLogic dilution data set (CNS). Four normalization methods, (1) LR (Linear Regression), (2) Loess (Loess Regression), (3) IR (Iterative Ranking) and (4) INR (Iterative Nonlinear Regression), are compared in terms of expression variance reduction. The normalization 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.

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:

$$\varepsilon_{\text{foldchange}} = \frac{1}{N}\sum_{i=1}^{N}(R_i - R_i^0)^2, \qquad (3)$$

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 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 $\varepsilon_{\text{foldchange}}$ when using INR than that using IR. The percentage of the genes with larger $\varepsilon_{\text{foldchange}}$ was then calculated for assessing the performance according to Criterion 2 [14].

Fig. (**9**) shows the comparison results of fold-change preservation on the first testing data set (Liver). The performances 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 $\varepsilon_{\text{foldchange}}$ than that using IR method, and 100% of genes having larger $\varepsilon_{\text{foldchange}}$ than that using INR method. Third, INR method gave the best performance in terms of fold-change preservation, only 16% of genes having larger $\varepsilon_{\text{foldchange}}$ than that using IR method. Note that the table of Fig. (**9**) was calculated with the "greater than" relation (i.e., the ">" relation). Therefore, the sum of (INR>IR: 16% for example) and (IR>INR: 78%) is 94%, which is not equal to 100%. This is because there are 6% of genes are of equal $\varepsilon_{\text{foldchange}}$ in Fig. (**9**) (i.e., the "=" relation). This applies to Figs. (**10**,**11**), therafter as well.

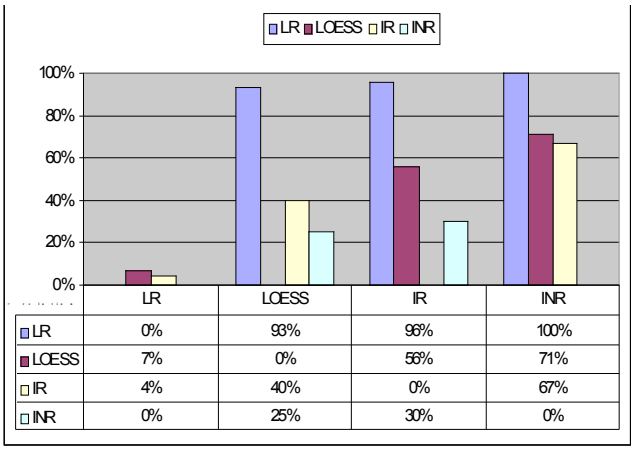


	LR	LOESS	IR	INR
LR	0%	100%	100%	100%
LOESS	0%	0%	98%	100%
IR	0%	4%	0%	78%
INR	0%	0%	16%	0%

Fig. (9). Fold-change comparison using GeneLogic dilution data set (Liver). Four normalization methods, (1) LR (Linear Regression), (2) Loess (Loess Regression), (3) IR (Iterative Ranking) and (4) INR (Iterative Nonlinear Regression), 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.

Fig. (**10**) shows the comparison results on the second testing data set (CNS). 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 $\varepsilon_{\text{foldchange}}$ than that using IR method (or Loess method).

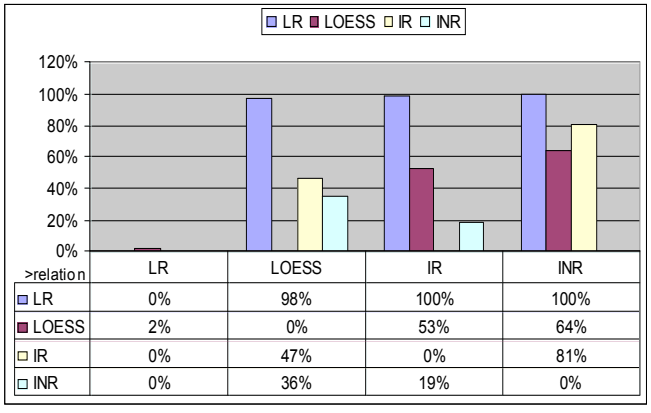
To further test the performance on the genes with small fold-changes, we constructed a third set with many small fold-changes (0.5< fold-change <2.0), 700 genes that were consistently detected as present across five low-intensity replicate arrays (5μg Liver cRNA) and 1400 from high-intensity replicate arrays (10μg Liver cRNA) were randomly selected. Fourteen 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 700 and 1400 genes selected. The expression measurements of the selected genes in each of the fourteen sets were then multiplied by 1.2, 0.83, 1.4, 0.71, 1.6, 0.63, 1.8, 0.56, 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.



	LR	LOESS	IR	INR
LR	0%	93%	96%	100%
LOESS	7%	0%	56%	71%
IR	4%	40%	0%	67%
INR	0%	25%	30%	0%

Fig. (10). Fold-change comparison using GeneLogic dilution data set (CNS). Four normalization methods, (1) LR (Linear Regression), (2) Loess (Loess Regression), (3) IR (Iterative Ranking) and (4) INR (Iterative Nonlinear Regression), 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.

Fig. (**11**) shows the comparison results on the third testing data set. Among all four normalization methods, LR method was again the worst one in terms of fold-change preservation. As seen from the figure, INR method continued to show the best performance in preserving fold-changes, specifically, only 19% (or 36%) of the genes having larger $\varepsilon_{\text{foldVchange}}$ than that using IR method (or Loess method). Note that compared to Fig. (**9**), we can see that adding in small fold-changes does worsen the performance although not significantly; in particular, the percentage of the genes using INR having larger $\varepsilon_{\text{foldVchange}}$ than using Loess method increases from 0% in Fig. (**9**) to 36% in Fig. (**11**).



>relation	LR	LOESS	IR	INR
LR	0%	98%	100%	100%
LOESS	2%	0%	53%	64%
IR	0%	47%	0%	81%
INR	0%	36%	19%	0%

Fig. (11). Fold-change comparison using GeneLogic dilution data set (Liver) with many small fold-changes (0.5< fold-change <2.0). Four normalization methods, (1) LR (Linear Regression), (2) Loess (Loess Regression), (3) IR (Iterative Ranking) and (4) INR (Iterative Nonlinear Regression), are compared in terms of fold change preservation. The table should be interpreted as in the following example: (INR, IR) = 19% means that with INR method, only 19% of the differentially expressed genes are of larger fold-change than that with IR method.

CONCLUSIONS

In this paper, we have reported a new method, INR, for microarray data normalization. The INR method features an iterative procedure for selecting IEGs and performing non-linear regression normalization. By cycling back and forth between the following two steps - (1) identifying control genes and (2) estimating regression functions, we can effectively identify the underlying IEGs for normalization. In particular, we have devised an efficient algorithm to identify the IEGs by gradually reducing the size of a sector-shaped region while moving the IEGs to be close to the 45-degree line in scatter plot.

We tested the INR method on two real microarray data sets – GeneLogic's array data set for dilution study and CNMC's microarray data set for muscular dystrophy study. The experimental results have demonstrated that an improved performance can be obtained using INR method 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.

We also compared the performance of INR method with other three widely adopted methods (i.e., LR, Loess and IR methods). The performance was evaluated based on the following two criteria: (1) expression variance 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.

ACKNOWLEDGEMENTS

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.

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*, vol. 286, pp. 531-7, 1999.

[2] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, 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," *Nat Med*, vol. 7, pp. 673-9, 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," *Nat Genet*, vol. 34, pp. 166-76, 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 U S A*, vol. 95, pp. 14863-8, 1998.

[5] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of

gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc Natl Acad Sci U S A*, vol. 96, pp. 2907-12, 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.*, vol. 11, pp. 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 U S A*, vol. 98, pp. 15149-54, 2001.

[8] J. Xuan, Y. Wang, Y. Dong, Y. Feng, B. Wang, J. Khan, M. Bakay, Z. Wang, L. Pachman, S. Winokur, Y.-W. Chen, R. Clarke, and E. Hoffman, "Gene selection for multiclass prediction by weighted Fisher criterion," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, pp. 1-15, 2007.

[9] M. Bilban, L. K. Buehler, S. Head, G. Desoye, and V. Quaranta, "Normalizing DNA microarray data," *Curr Issues Mol Biol*, vol. 4, pp. 57-64, 2002.

[10] A. Hartemink, D. Gifford, T. Jaakkola, and R. Young, "Maximum likelihood estimation of optimal scaling factors for expression array normalization," presented at SPIE BiOS 2001: International Biomedical Optics Symposium San Jose, CA, 2001.

[11] G. C. Tseng, M. K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong, "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects," *Nucleic Acids Res*, vol. 29, pp. 2549-57, 2001.

[12] Affymetrix, "Affymetrix technical note: Statistical algorithms description document," in *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 slide systematic variation," *Nucleic Acids Res*, vol. 30, pp. e15, 2002.

[14] E. E. Schadt, C. Li, B. Ellis, and W. H. Wong, "Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data," *J Cell Biochem Suppl*, vol. Suppl 37, pp. 120-5, 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*, vol. 19, pp. 185-93, 2003.

[16] J. Quackenbush, "Microarray data normalization and transformation," *Nat Genet*, vol. 32 Suppl, pp. 496-501, 2002.

[17] E. Camerer, E. Gjernes, M. Wiiger, S. Pringle, and H. Prydz, "Binding of factor VIIa to tissue factor on keratinocytes induces gene expression," *J Biol Chem*, vol. 275, pp. 6580-5, 2000.

[18] 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*, vol. 61, pp. 5974-8, 2001.

[19] 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*, vol. 2, pp. RESEARCH0055, 2001.

[20] Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke, "Iterative normalization of cDNA microarray data," *IEEE Trans. on Info. Tech. in Biomedicine*, vol. 6, pp. 29-37, 2002.

[21] G. Wahba, *Spline methods for observational data*: Philadelphia: SIAM, 1990.

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

[23] GeneLogic, "Dilution/mixture datasets," in *http://www.genelogic.com*, 2002.

[24] 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. T. Winokur, L. M. 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*, vol. 129, pp. 996-1013, 2006.

Received: June 5, 2007 Revised: June 27, 2007 Accepted: August 8, 2007