Pacific Symposium on Biocomputing 6:496-507 (2001)

# ON REPORTING FOLD DIFFERENCES

C.L. TSIEN
*Massachusetts Institute of Technology and Harvard Medical School*
*545 Technology Square, NE43-420, Cambridge, MA 02139, USA*

T.A. LIBERMANN, X. GU
*New England Baptist Bone and Joint Institute, Beth Israel Deaconess Medical Center, and Harvard Medical School, 4 Blackfan Circle, Boston, MA 02115, USA*

I.S. KOHANE
*Children's Hospital Informatics Program and Harvard Medical School*
*300 Longwood Avenue, Boston, MA 02115, USA*

As we enter an age in which genomics and bioinformatics make possible the discovery of new knowledge about the biological characteristics of an organism, it is critical that we attempt to report newly discovered "significant" phenotypes only when they are actually of significance. With the relative youth of genome-scale gene expression technologies, how to make such distinctions has yet to be better defined. We present a "mask technology" by which to filter out those levels of gene expression that fall within the noise of the experimental techniques being employed. Conversely, our technique can lend validation to significant fold differences in expression level even when the fold value may appear quite small (e.g. 1.3). Given array-organized expression level results from a pair of identical experiments, our ID Mask Tool enables the automated creation of a two-dimensional "region of insignificance" that can then be used with subsequent data analyses. Fundamentally, this should enable researchers to report on findings that are more likely to be in nature truly meaningful. Moreover, this can prevent major investments of time, energy, and biological resources into the pursuit of candidate genes that represent false positives.

## 1 Introduction

As we enter one of the most exciting times in the history of science, in which genomics and bioinformatics are coming together to make possible the discovery of new knowledge about living organisms at their molecular level, it is imperative that we avoid discovery of "truths" that are not so. While the temptation to plunge into tracing out metabolic pathways, cellular interactions, or genetic regulatory circuits—especially now that we have technologies allowing genome-wide study of RNA expression—is very strong, we must pause long enough to consider how best to report our results such that they may be meaningful. Specifically, for microarray-based expression technologies, whether they are glass microarrays, nylon membranes, or other formats, we need to better understand how to distinguish significant fold difference values from those that fall within the noise level of the experiment at hand.

Francis Collins rightfully speculates about the large impact that microarray technology is likely to have, yet reminds us of the "many critically important questions about this new field that are yet unaddressed" [1]. Some have criticized array-based methods for not being model-based, or hypothesis-driven, while others support that the exploratory nature can lead to new hypotheses that then can be tested in the laboratory [2]. Especially because such hypothesis testing of candidate genes, cell-cell interactions, or pathways requires a major investment of time, energy, and biological resources, an important challenge is understanding how to better recognize false-positive results.

We present a "mask technology" by which to filter out those levels of gene expression that fall within the noise of the experimental techniques being employed. Conversely, our technique can lend validation to the significance of fold differences in expression level even when the fold value may appear quite small. Our work is based on the notion that gene expression measurements ought to be repeatable. Fold differences for each corresponding pair of genes in a pair of "identical" experiments should therefore be equal to unity. Identical experiments are ones in which the operating conditions, cell lines, culture media, incubation time, and so forth are controlled to be the same. We first explore whether this is the case by examining several pairs of identical experiments. We then develop the ID Mask Tool, which enables the automated creation of a two-dimensional "region of insignificance" that can be used with subsequent data analyses.

## 2 Materials and Methods

### *2.1 Data Collection*

The data for this study were collected to evaluate the use of microarray technology for detection of ESE-1 target genes after transient transfection into different cell lines. We hypothesized that a transfection efficiency of greater than 70-80% should be sufficient to detect differences in gene expression between two samples. We first determined the transfection efficiency of various cell lines using a green fluorescent protein (GFP) expression vector. Four of the cell lines tested (HT1080, 293, MCF-7, and MG-63) conformed to the criteria set by us. Total RNA was isolated from MCF-7 human breast cancer cells and MG-63 human osteosarcoma cells transiently transfected with an ESE-1 expression vector 20 and 24 hours after transfection. Experiments were performed in duplicates in order to distinguish, from gene expression, differences due to "biological noise." Specifically, six pairs of these duplicated experiments served as the source of the data that we subsequently used to develop the identity mask methodology. The ESE-1 expression vector also

expressed GFP, which enabled us to confirm transfection efficiencies for each experiment. $^{32}$P-labeled cDNA probes reverse-transcribed from these RNAs were hybridized to the Atlas Human cDNA Expression Arrays from Clontech (Clontech Laboratories, Inc., Palo Alto, CA) [3]. Each of these Atlas Arrays (Human 1.2 I, Human Cancer) is a nylon membrane on which approximately 1200 human cDNAs have been immobilized. The hybridization results were analyzed with the software provided by Clontech by normalizing to the signals obtained from housekeeping gene controls on the same array as well as by global normalization. The microarray experiments were validated by RT/PCR using the same RNAs.

### *2.2 Data Analysis and Mask Creation*

We developed the ID Mask Tool, a custom-designed computer program written in the C language, to perform mask creation. The ID Mask Tool takes as input two spreadsheet files corresponding to two identical experiments, along with two user customizable parameters to be discussed below. It returns as output an "identity mask," or ID Mask, specifically for those two experiments.

Each spreadsheet contains the names of several hundred genes and their corresponding brightness intensity levels (as assessed by hybridization of the probe of interest). Only genes present in both files are further considered. For each of these genes, we calculate a "fold difference," the ratio of the intensity in file 2 to the intensity in file 1 for a given gene. All fold values are then sorted based on the corresponding intensity values of the set of genes in the first spreadsheet file. Two parameters are used for creation of each identity mask: intensity range (or sliding window) size, plus either scale value or number of standard deviations. These are used to calculate the ID Mask borders and can be experimented with for better results.

Two methods are then explored for creating identity masks. Method 1 relies on segmental calculation of standard deviations. A "data point" refers to an ($x$, $y$) pairing in which $x$ is an intensity value from the first spreadsheet file and $y$ is its corresponding fold difference value (calculated as above). Using all data points in a given sliding window of intensity values (e.g., from intensity level 1001 to 2000), the standard deviation of the fold values is calculated. The average of the intensity values within that window is then paired with a fold value equal to the average fold value within that window plus the number of standard deviations specified by the user. This new pair becomes a candidate "upper mask border" point. Similarly, a candidate "lower mask border" point is created by pairing the average intensity value of that window with the average fold value minus the number of standard deviations specified by the user. Each successive group of data points in each

sliding window of intensity values (e.g., all points from 2001 to 3000, then all points from 3001 to 4000, etc.) likewise gives rise to candidate mask border points.

The set of (*intensity value, fold value*) pairs comprising the candidate upper mask border points is then fit to a line using least-squares linear regression. This line defines the upper mask border. Similarly, linear regression is used to find the lower mask border from the set of calculated candidate lower mask border points. If one of the derived mask borders fits poorly (based upon relationship to original data points), the "reciprocal reflection" of the other mask border can serve in its place. This simply means that each *(x, y)* point on the good-fit (linear) border gives rise to a point (*x, 1/y)* to create the reciprocal reflection border. (See Figures 1 through 6 for examples of mask borders. Figures 2—5 show ID Masks each consisting of one linear regression border and one border derived by taking the reciprocal values of that linear regression border.) The region between these borders represents the "identity" region of insignificant fold differences (i.e., noise).

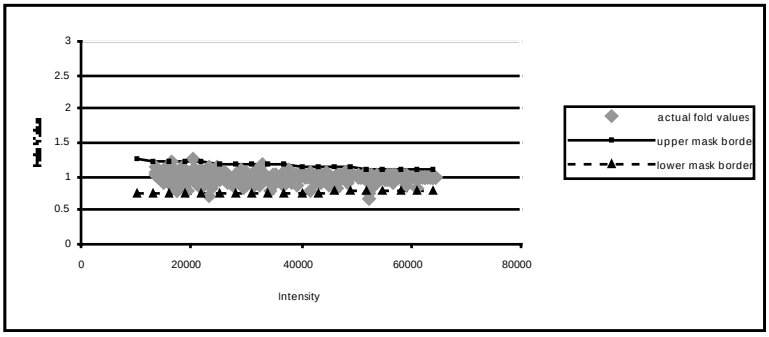


**Figure 1:** Identity mask for Experiment A. Method 2 with parameters 9000 for intensity sliding window size and 0.975 for scale resulted in the lowest percentage of original data points lying outside of the mask region (0.7%).

Method 2 for creating an identity mask is similar to Method 1 except that candidate mask border points are derived from maximal (and minimal) points in each intensity window rather than from standard deviation calculations. Specifically, amongst all data points in a given window of intensity values, the point with the greatest fold value is chosen. This is repeated for each successive window of intensity values. These fold values can also be scaled before use in linear regression to find the upper mask border. The lower mask border is analogously derived from the smallest fold values.

Once the ID Mask has been derived, all original data points are checked for inclusion or exclusion in the identity mask region. The percentage of data points lying outside of the mask region is reported.

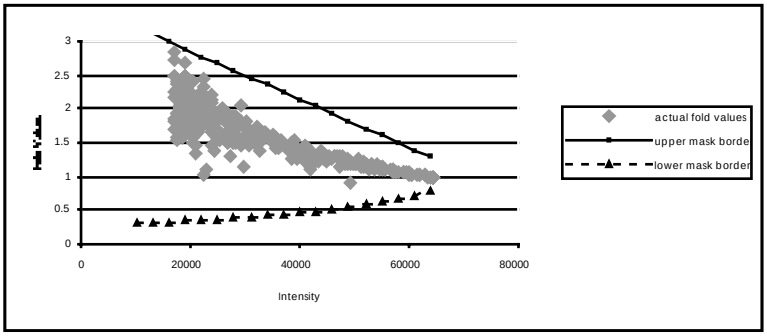


**Figure 2:** Identity mask for Experiment B. Method 1 with parameters 9000 for intensity window size and standard deviation of 3 resulted in the lowest percentage of original data points lying outside of the mask region (1.7%).

**Table 1:** Numbers of genes present in each of the experiment pairs, along with the number of genes common to both files in each pair.

| | **# Genes in 1st File** | **# Genes in 2nd File** | **# Genes in Both** |
|---|---|---|---|
| **Expt A** | 563 | 559 | 550 |
| **Expt B** | 292 | 516 | 291 |
| **Expt C** | 244 | 401 | 244 |
| **Expt D** | 339 | 518 | 326 |
| **Expt E** | 365 | 397 | 344 |
| **Expt F** | 233 | 226 | 180 |

## 3 Results

Six pairs of experiments were performed with Clontech nylon membrane filters and tumor cell lines as described in the Methods section, resulting in twelve spreadsheet files of genes and their corresponding expression intensity values. The ID Mask Tool was used to perform all mask creation experiments as well as basic data analysis. Table 1 displays the number of genes present in each of the file pairs, along with the number of genes common to both files in each pair.

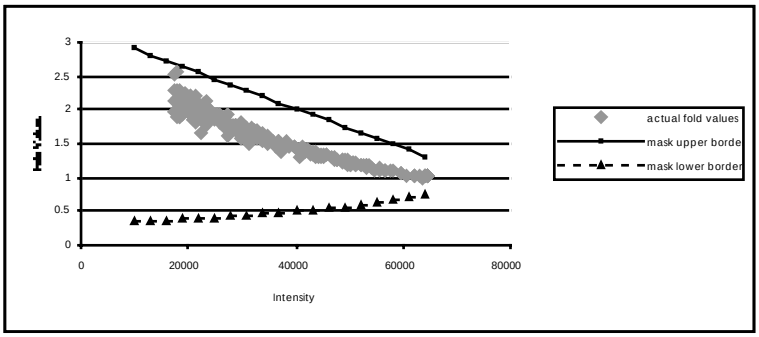


**Figure 3:** Identity mask for Experiment C. Method 1 with parameters 9000 for intensity window size and standard deviation of 3 resulted in the lowest percentage of original data points lying outside of the mask region (2.0%).

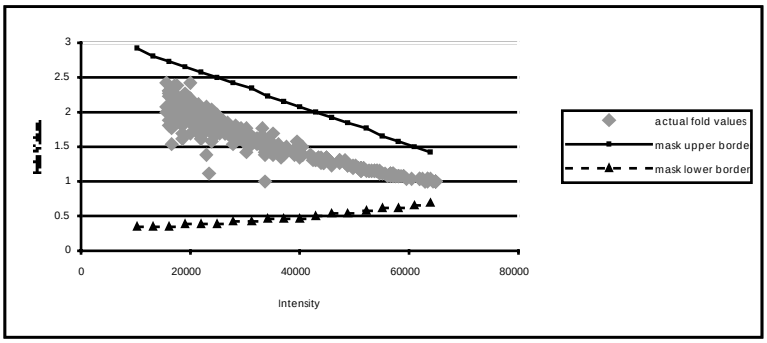


**Figure 4:** Identity mask for Experiment D. Method 1 with parameters 9000 for intensity window size and standard deviation of 3 resulted in the lowest percentage of original data points lying outside of the mask region (1.5%).

For both Methods 1 and 2 of ID Mask creation, sliding windows of size 1000, 5000, and 9000 on the intensity value axis were chosen for experimentation. Only when calculations were not possible with one of these window sizes (e.g., due to division by zero) was an alternative window size chosen. For Method 1, the number of standard deviations (for calculation of candidate mask border points) was chosen to be 2.5 and 3. For Method 2, the scale factor was chosen to be 0.975 and 1.0.

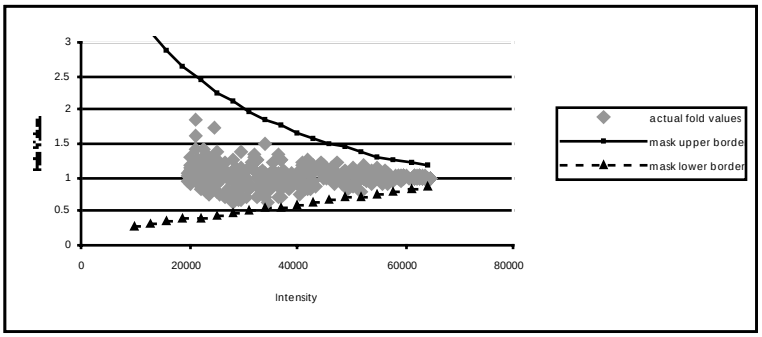


**Figure 5:** Identity mask for Experiment E. Method 1 with parameters 5000 for intensity window size and standard deviation of 3 resulted in the lowest percentage of original data points lying outside of the mask region (0.9%).

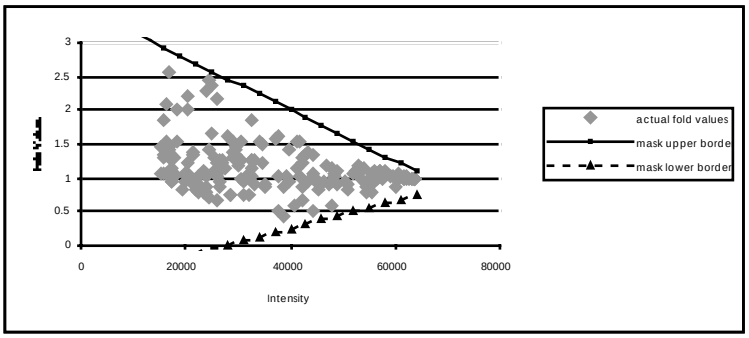


**Figure 6:** Identity mask for Experiment F. Method 1 with parameters 9000 for intensity window size and standard deviation of 3 resulted in the lowest percentage of original data points lying outside of the mask region (1.7%).

Twelve candidate identity masks were created for each pair of experiments (2 Methods, times 3 intensity window sizes, times 2 scale or standard deviation factors). For each pair of experiments, the ID Mask Tool selected the mask with the lowest percentage of original data points lying outside of the mask region. Figures 1 through 6 show each selected identity mask along with a scatter plot of the original *(intensity value, fold value)* data points for each pair of experiments. Tables 2 and 3 list the percentages of original data points lying outside of the mask region for each of the 12 candidate masks derived for each experiment pair.

**Table 2:** Each pair of identical experiments gave rise to 12 candidate ID Masks. Six of these twelve were derived by Method 1 (three with standard deviation of 3 and three with standard deviation of 2.5). The other six were derived by Method 2 (three with scale 1.00 and three with scale 0.975). Shown are the percentages of original data points lying outside of the mask region for each of the 12 candidate ID Masks derived for each of Experiments A—C. [σ = standard deviation; intensity range (window) size of 2000 instead of 1000 is used in Experiments B and C for the Method 1 trials.]

| | **Expt A** | **A** | **A** | **Expt B** | **B** | **B** | **Expt C** | **C** | **C** |
|---|---|---|---|---|---|---|---|---|---|
| range size | 1000 | 5000 | 9000 | 2000; 1000 | 5000 | 9000 | 2000; 1000 | 5000 | 9000 |
| σ = 3 | 3.1 | 2.7 | 2.2 | 93.2 | 80.5 | 1.7 | 99.6 | 100.0 | 2.0 |
| σ = 2.5 | 11.1 | 3.8 | 3.3 | 97.3 | 97.6 | 2.7 | 100.0 | 100.0 | 2.4 |
| Scale 1.00 | 19.2 | 6.2 | 0.7 | 100.0 | 99.0 | 2.4 | 100.0 | 100.0 | 2.4 |
| Scale 0.975 | 19.2 | 6.2 | 0.7 | 100.0 | 99.3 | 2.4 | 100.0 | 100.0 | 2.9 |

**Table 3:** Percentages of original data points lying outside of the mask region for each of the 12 candidate ID Masks derived for Experiments D—F. (See caption in Table 2 for further details.) [σ = standard deviation; intensity range (window) size of 3000 instead of 1000 is used in Experiment D for Method 1 trials, while range (window) size of 2000 instead of 1000 is used in Experiment F for Method 1 trials.]

| | **Expt D** | **D** | **D** | **Expt E** | **E** | **E** | **Expt F** | **F** | **F** |
|---|---|---|---|---|---|---|---|---|---|
| range size | 3000; 1000 | 5000 | 9000 | 1000 | 5000 | 9000 | 2000; 1000 | 5000 | 9000 |
| σ = 3 | 17.1 | 88.4 | 1.5 | 0.9 | 0.9 | 0.9 | 5.5 | 6.1 | 1.7 |
| σ = 2.5 | 24.5 | 99.4 | 1.5 | 2.6 | 2.0 | 2.9 | 7.2 | 6.6 | 2.8 |
| Scale 1.00 | 100.0 | 99.7 | 2.4 | 43.8 | 18.6 | 19.7 | 53.0 | 9.9 | 10.5 |
| Scale 0.975 | 100.0 | 99.7 | 3.4 | 44.3 | 20.0 | 20.0 | 54.7 | 9.9 | 11.6 |

## 4 Discussion

DNA microarrays clearly are making a large impact on the way we approach problems in molecular biology and genomics. These devices are enabling the genome-wide study of expression in *Escherichia coli* K-12, for example [4]. Others are using DNA microarrays in the study of B-cell lymphomas [5], growth control genes [6], and aging [7]. Some researchers are focusing on developing new [8] or using existing [9] clustering techniques to facilitate the analysis of all the data made available by this relatively new technology. Few, however, have focused specifically on studying the properties of these array data to better understand how to distinguish significant from insignificant "findings."

One way we might be able to better discern meaningful discoveries from the rest is by applying an identity mask technology, such as the one we have presented. Our experiments show that greater amounts of biological noise are present at lower gene expression levels. Thus, there is no magical absolute cut-off for a meaningful fold value. There does appear to exist, however, a "mask of insignificant values," outside of which the fold values are more likely to represent true significance. In Figure 6, for example, a fold difference of 1.5 may be meaningful at an intensity level of 60,000, while a fold difference of 2.5 may be insignificant at an intensity level of 20,000. This result is in stark contrast to a study by Incyte Pharmaceuticals [11], in which they conclude: "any elements with observed ratios greater than or equal to 1.8 should be deemed differentially expressed." A brief glance at the microarray-related literature will quickly confirm that others are also reporting particular fold difference values, such as 1.8, as significant [7]. We argue, however, that the significance of a fold change depends upon the intensity value; genes that are expressed at low levels and hence have weak intensity signals need to show a much greater fold difference than highly expressed genes.

Some have proposed simple statistical tests to determine whether fold differences are significant; t-tests, for example, are included in the GeneSpring software package (Silicon Genetics, San Carlos, CA). Lee *et al.* propose a statistical method using normal distributions and posterior probabilities to determine the likelihood that a gene is truly expressed in a tissue sample [12]. Methods like these are no doubt important; used alone, however, they may under-emphasize the correlation between fold values and intensity values. Future efforts might explore how to best use statistical validation techniques in conjunction with the identity mask method.

While our study used Clontech filters, the general techniques presented for understanding identity masks of insignificance apply to all different types of expression arrays. Both nylon membrane and glass slide array techniques have their

individual advantages. Nylon membrane arrays have sensitive detection using hybridized $^{32}P$ probes. Glass microarrays have high-resolution fluorescent detection, dual labeling for hybridizing two probes on a single array, and ease in automated handling of slides [3]. Richmond *et al.* compared hybridization of radioactive cDNAs to spot blots on nylon membranes with fluorescence-based hybridization to glass microarrays; they found both methods to be reliable and reproducible [4]. Chen describes a colorimetry detection system for use with nylon membranes [13].

Regardless of the specific array format employed, it seems clear that a custom-derived identity mask is one method that could help improve appropriate reporting of fold difference results. Future work should include an exploration of fitting curves rather than lines for the mask borders. The upper mask border in Figure 2, for example, may benefit from a fitted curve, or at least a piecewise linear model.

An alternative method for mask creation might be to always calculate fold differences greater than 1 by simply swapping the order of individual intensity values whenever the fold value is less than 1. Only the upper mask border would then need to be created. (The lower mask border would be the unity fold difference line.)

It is not clear why there were some large differences between the numbers of genes detected in the experiment pairs of Experiments B, C, and D. These may have been due to experimental error or biological noise. Interestingly, the identity masks for these three also do not fit as nicely as those for Experiments A, E, and F.

While we have selected from amongst the candidate identity masks those with the lowest percentages of points outside the mask region, future work might consider refining the mask fit to purposely exclude approximately 5% of the data points. This could be likened to $p < 0.05$, in which 5% of the time, we may inadvertently report a result as significant even though it is not. A potential benefit is a closer overall mask fit and therefore less likelihood to call a significant finding insignificant.

In only one out of the six pairs of experiments did Method 2 (scaling values) perform better than Method 1 (standard deviations). This is possibly due to the mathematical basis upon which standard deviations are calculated, making them in general more robust and accurate. One way in which scaling actual data points can fail is when there exist outliers. Another is with the choice of too small an intensity window size. This can lead to a sort of “overfitting” problem; our group of candidate “maximum” points from which to derive the upper mask border may then contain several non-maximum values. In Tables 2 and 3, there is a definite trend of worsening mask fit as one decreases the intensity range (window) size from 9000 to 1000. It is likely that in most applications, Method 1 may be more suitable.

Our aim has been to provide a foundation for evaluating fold values. The ultimate goal is to find truly significant fold differences when performing "treatment versus control" comparisons. Analyses of those types of comparisons will likely further our understanding of the masking technique as well. Especially because we recognize the use of DNA microarrays as a method by which to explore the genome in a model-independent fashion [10], it is imperative that we have a basis for judging exploratory findings as being important or simply "in the noise." Candidate genes found through exploration can lead to investment of significant resources; we need to avoid such pursuits of false positive findings.

## Acknowledgments

The authors would like to thank Atul Butte, M.D. for assistance with references, and the members of Towia Libermann's laboratory for performing the hybridization experiments.

## References

1. F.S. Collins, "Microarrays and macroconsequences" *Nature Genetics* Suppl. **21**, 2 (1999)
2. "The chip challenge" *Nature Genetics* **21**, 61-62 (1999)
3. Clontech web page at http://www.clontech.com/about/index.html
4. C.S. Richmond, J.D. Glasner, R. Mau, et al., "Genome-wide expression profiling in Echerichia coli K-12" *Nucleic Acids Research* **27**, 3821-3835 (1999)
5. A.A. Alizadeh, M.B. Eisen, R.E. Davis, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling" *Nature* **403**, 503-511 (2000)
6. F. Bertucci, S.V. Hulst, K. Bernard, et al., "Expression scanning of an array of growth control genes in human tumor cell lines" *Oncogene* **18**, 3905-3912 (1999)
7. C.-K. Lee, R.G. Klopp, R. Weindruch, et al., "Gene expression profile of aging and its retardation by caloric restriction" *Science* **285**, 1390-1393 (1999)
8. M.B. Eisen, P.T. Spellman, P.O. Brown, et al., "Cluster analysis and display of genome-wide expression patterns" *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868 (1998)
9. A.J. Butte and I.S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements" *Proceedings of the Pacific Symposium on Biocomputing* (2000)

10. P.O. Brown and D. Botstein, “Exploring the new world of the genome with DNA microarrays” *Nature Genetics* Suppl. **21**, 33-37 (1999)
11. “GEM microarray reproducibility study” Incyte Technical Survey, Incyte Pharmaceuticals, Inc. (1999)
12. M.T. Lee, F.C. Kuo, G.A. Whitmore, et al., “Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations” *PNAS* **97**, 9834-9839 (2000)
13. J.J. Chen, R. Wu, P.C. Yang, et al., “Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection” *Genomics* **51**, 313-324 (1998)