Journal of Molecular Diagnostics, Vol. 5, No. 2, May 2003
Copyright © American Society for Investigative Pathology
and the Association for Molecular Pathology

Analysis of Microarray Data Using Z Score Transformation

Chris Cheadle,* Marquis P. Vawter,†‡
William J. Freed,† and Kevin G. Becker*

From the DNA Array Unit, Research Resources Branch, National Institute on Aging, National Institutes of Health, Baltimore, Maryland; the Cellular Neurobiology Branch,† National Institute on Drug Abuse, Baltimore, Maryland; and the Department of Psychiatry and Human Behavior,‡ University of California, College of Medicine, Irvine, California*

High-throughput cDNA microarray technology allows for the simultaneous analysis of gene expression levels for thousands of genes and as such, rapid, relatively simple methods are needed to store, analyze, and cross-compare basic microarray data. The application of a classical method of data normalization, Z score transformation, provides a way of standardizing data across a wide range of experiments and allows the comparison of microarray data independent of the original hybridization intensities. Data normalized by Z score transformation can be used directly in the calculation of significant changes in gene expression between different samples and conditions. We used Z scores to compare several different methods for predicting significant changes in gene expression including fold changes, Z ratios, Z and t statistical tests. We conclude that the Z score transformation normalization method accompanied by either Z ratios or Z tests for significance estimates offers a useful method for the basic analysis of microarray data. The results provided by these methods can be as rigorous and are no more arbitrary than other test methods, and, in addition, they have the advantage that they can be easily adapted to standard spreadsheet programs. ***(J Mol Diagn 2003, 5:73–81)***

cDNA microarray technologies are rapidly being applied in biology and medicine.[1] It is hoped that mining of microarray datasets will lead to the discovery of pathways of common and unique gene expression among different cells, tissues, and disease states.[2] The use of cDNA arrays among individual investigators, between laboratories, and across disciplines has necessitated the reporting of expression data that can be rapidly compared and can be easily archived.

Before comparing the microarray results from multiple experiments the results from individual experiments must somehow be normalized with respect to each other to account for experimental variation in RNA amounts, specific activity of cDNA labels, and standard handling errors. Failure to properly normalize data used in microarray comparisons runs a high risk of skewing comparison results and reduces the credibility of individual gene change measurements. One of the most common ways in which microarray data are normalized is to assume that the majority of gene expression is relatively constant between experiments and that this constant population can serve as the basis for a general approach to normalization. Empirical observation, in almost all cases, continues to support this underlying assumption used in population normalizations. Occasionally the use of population normalization may be contraindicated when, for example, a highly restricted subset of genes is used to measure a highly dynamic biological condition (eg, a small focused array used to study embryonic development). In this case an alternative normalization method such as spiking of internal references[3] should be considered. Clearly, the experimental design must be carefully evaluated before the selection of an appropriate normalization technique.

One basic method of population normalization is global normalization,[4] which calculates the mean or median of the signal intensities of each individual experimental dataset and then calculates the mean of the means (or grand mean) for all of the included experiments. Each individual data set is then mathematically adjusted such that the mean of that dataset equals the calculated grand mean. This method is conceptually simple, but when working with datasets having large differences in signal intensity, the data can be inordinately influenced by the presence of outlier data distortions. In addition, and equally as problematic from a high throughput standpoint, each time another experiment is added to the experimental comparisons, the collective or grand mean must be recalculated, and all of the experimental datasets readjusted.

Supported by the William Lion Penzner Foundation (MPV).

C. C. and M.P.V. contributed equally to this work.

Accepted for publication December 16, 2002.

Address reprint requests to Kevin G. Becker, Ph.D., Chief-DNA Array Unit, Research Resources Branch, National Institute on Aging, NIH, 333 Cassell Drive, Baltimore, MD 21221-6825. E-mail: beckerk@grc.nia.nih.gov.

Here we describe the normalization and standardization of cDNA microarray intensity values within datasets by Z score transformation and the subsequent use of the transformed data to compare multiple experiments. The Z score transformation procedure for normalizing data is a familiar statistical method in both neuroimaging[5] and psychological studies,[6,7] among others. Recently, Z score transformation statistics have been used in comparing experimental and control group gene expression[8–10] differences by microarray. Z score transformation methods have also been incorporated into the latest version of the public access MAExplorer (supplied by Peter Lemkin of the National Cancer Institute) microarray bioinformatics tool.[11]

The Z score transformation approach for microarrays corrects data internally within a single hybridization and hybridization values for individual genes are expressed as a unit of SD from the normalized mean of zero. Correction is done before sample-to-sample comparison, and is therefore comparison-independent. Comparisons across samples or across experiments are then performed on equivalently transformed data, and changes in gene expression are expressed as differences between Z scores (Z ratios) or by using a statistical test such as the two-sample-for-means Z test.[12] Using this approach, gene expression data derived from different microarray studies becomes comparable across experiments and across laboratories.

In this paper we have compared differences in gene expression using the traditional fold-ratios (arrived at by global normalization) and Z ratios (calculated from Z scores). In addition, we compared significance levels derived from several different statistical methods including Z ratios, Z tests, and *t*-tests, combined with permutation analysis using Significance Analysis of Microarrays (SAM) software from Stanford University labs[13]. We have chosen a dataset that is relatively simple to reduce the number of parameters to be considered in making the comparisons between global and Z score transformations, as well as the comparisons between statistical tests. The Z score transformation process is easily extended to more complex datasets.

Materials and Methods

Cells and Tissue Cultures

Freshly purified human peripheral blood T (PBT) cells obtained from three different donors were greater than 95% $CD3^+$ cells. Human PBT cells as well as Jurkat human T cells were cultured in RPMI 1640 with 10% fetal bovine serum, 100 U/ml penicillin, 100 μg/ml streptomycin and 2 mmol/L glutamine. The PBTs were allowed to stabilize overnight before testing. PMA was used at 10 ng/ml, ionomycin at 1 μg/ml, and anti-CD28 monoclonal antibody (kindly provided by Dr. Carl H. June) at 100 ng/ml.

RNA Purification

Total cellular RNA was extracted after 2 hours of stimulation directly from T cells in three conditions: control T cells (Ct), T cells following stimulation by phorbol myristate plus ionomycin (PMA+I), and T cells following phorbol myristate plus anti-CD28 antibody (PMA + 28). The total RNA was extracted in the flasks using a one-step guanidine thiocyanate/phenol method[14] followed by sequential ethanol precipitations. The concentration and quality of the RNA were assessed by spectrophotometry and by agarose gel electrophoreses. RNA samples were stored at −80°C until used.

RNA Labeling

RNA samples were radiolabeled and hybridized according to protocols described in http://www.grc.nia.nih.gov/branches/rrb/dna.htm. For probe preparation (radiolabeling of total RNA with [^{33}P]dCTP), 5 μg of total RNA for each sample was radiolabeled in a reverse-transcription (RT) reaction. RNA was annealed, in 16 μl H_2O, with 1 μg of 24-mer poly(dT) primer (Research Genetics, Huntsville, AL), by heating at 65°C for 10 minutes and cooling on ice for 2 minutes. The RT reaction was performed by adding 8 μl of 5X first-strand RT buffer (Life Technologies, Rockville, MD), 4 μl of 20 mmol/L dNTPs minus dCTP) (Pharmacia, Piscataway, NJ), 4 μl of 0.1 mol/L DTT, 40 U of RNAseOUT (Life Technologies), 6 μl of 3000 Ci/mmol α[^{33}P]dCTP (ICN Biomedicals, Costa Mesa, CA) to the RNA/primer mixture to a final volume of 40 μl. Two μl (400 U) of Superscript II reverse transcriptase (Life Technologies) was then added, and the sample was incubated for 30 minutes at 42°C followed by additional 2 μl of Superscript II reverse transcriptase and another 30 minutes of incubation. The reaction was stopped by the addition of 5 μl of 0.5 mol/L EDTA. The samples were incubated at 65°C for 30 minutes after addition of 10 μl of 0.1 mol/L NaOH to hydrolyze and remove RNA. The samples were pH-neutralized by the addition of 45 μl of 0.5 mol/L Tris (pH 8.0) and purified using Bio-Rad 6 purification columns (Bio-Rad, Hercules, CA).

Microarray Construction and Use

Microarray construction and hybridization were performed as previously described.[15] Briefly, NIA-Immunoarrays, which consist of 1132 genes printed on Nytran + Supercharge nylon membranes (Schleicher & Schuell) in duplicate, were hybridized with α[^{33}P]dCTP-labeled cDNA probes overnight at 50°C in 4 ml of hybridization solution. Hybridized arrays were rinsed in 50 ml of 2X SSC and 1% SDS twice at 55°C followed by 1–2 times of washing in 2X SSC and 0.1% SDS at 55°C for 15 minutes each. The microarrays were exposed to phosphorimager screens for 1 to 3 days. The screens were then scanned in a Molecular Dynamics STORM PhosphorImager (Molecular Dynamics, Sunnyvale, CA) at 50 μm resolution. ImageQuant software (Molecular Dynamics) was used to convert the hybridization signals on the image into raw intensity values, and the data thus generated was transferred into Microsoft Excel spreadsheets, pre-designed to associate the ImageQuant data format to the correct gene identities.

Global Normalization

Raw intensity data for each experiment was normalized by first calculating the average intensity for each individual dataset, and then calculating the average of the averages. This grand average was used as the basis for the computation of normalization factors that were subsequently applied to each experiment. The average of all normalized data thereafter equaled the grand average.

Z *Score Transformation*

Raw intensity data for each experiment is $\log_{10}$ transformed and then used for the calculation of Z scores. Z scores are calculated by subtracting the overall average gene intensity (within a single experiment) from the raw intensity data for each gene, and dividing that result by the SD of all of the measured intensities, according to the formula:

$$\text{Z score} = (\text{intensity}_{G} - \text{mean intensity}_{G1...Gn})/\text{SD}_{G1...Gn}$$

where G is any gene on the microarray and G1. . . Gn represent the aggregate measure of all of the genes.

Estimate of Significant Changes in Gene Expression

Traditional Ratio

Traditional ratio calculations of significant changes in gene expression derived from globally normalized data are performed by simply computing the ratio of the average of all of the measurements from one condition or sample to another.[4] Significance is customarily assigned to genes whose ratio is greater or equal to 2.0 or less than or equal to 0.5.

Z *Ratios*

Z score values are used as the data basis in all calculations of changes in gene expression including Z ratios, Z tests, and SAM analysis. Z ratios are calculated by taking the difference between the averages of the observed gene Z scores and dividing by the SD of all of the differences for that particular comparison:

$$\text{Z ratio} = [(\text{Z score}_{\text{G1ave}})\text{Exp} - (\text{Z score}_{\text{G1ave}})_{\text{Con}}] / \text{SD of Z score differencesG1...Gn}$$

where G_1 represents the average Z score for any particular gene being tested under multiple experimental conditions (in this case, experimental versus control) and G1. . . Gn represents the aggregate measure of all of the genes. Calculated Z ratios have the advantage that they can be used in multiple comparisons without further reference to the individual conditional standard deviations by which they were derived. A Z ratio of ± 1.96 is inferred as significant ($P < 0.05$), although empirical observation has shown consistent results with Z ratio values of ± 1.50 or greater.

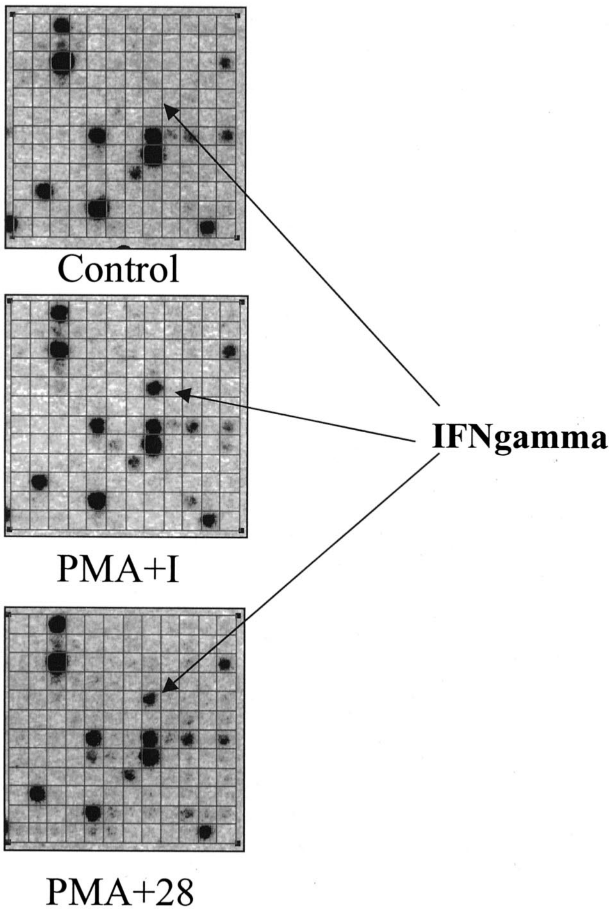


Figure 1. A portion of the hybridization images of three NIA-Immunoarray filters hybridized to radiolabeled total RNA from three different biological conditions (untreated T cells, control; PMA plus ionomycin, PMA+I; and PMA plus anti-CD28, PMA + 28). The grid used for quantitation is shown superimposed on each image. The increase in gene expression of interferon-γ (IFNgamma) is shown.

Z *Test*

An alternative method for calculating significant changes in gene expression, which maximizes the power of replicates and takes into account variation between replicates on a gene by gene basis, is the two-sample-for-means Z test.[12] The formula for this statistical test is as follows:

$$\text{Z test} = [(\text{Z score}_{\text{G1ave}})_{\text{Exp}} - (\text{Z score}_{\text{G1ave}})_{\text{Con}}] / \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

where G_1 represents the average Z score for any particular gene being tested under multiple experimental conditions (in this case, experimental versus control). The mean difference is corrected by the SE for the difference between means where σ_2 is the SD of repeated hybridization intensity measurements (expressed as Z scores) for either condition 1 or condition 2, and *n* equals the number of repeated measurements for either condition 1

Table 1. Comparison of Data Types: Original, Normalized, and **Z**-Transformation

	Original data			Normalized data			Fold change		Log data			Z score data			Z differences		Z ratios	
Gene	Ct	PI	P28	Ct	PI	P28	PI/Ct	P28/ Ct	Ct	PI	P28	Ct	PI	P28	PI-Ct	P28- Ct	PI- Ct	P28- Ct
IFNG	433	5964	2733	401	6692	2654	16.69	6.62	2.63	3.78	3.42	0.13	3.33	2.41	3.20	2.28	6.63	4.46
SCYA4	186	1563	1633	172	1754	1586	10.19	9.22	2.26	3.17	3.18	−0.87	1.80	1.57	2.67	2.44	5.53	4.78
MYC	286	2339	1912	265	2625	1856	9.90	7.00	2.45	3.37	3.28	−0.35	2.29	1.92	2.65	2.27	5.48	4.45
VIM	1423	7563	8731	1319	8486	8478	6.43	6.43	3.12	3.88	3.94	1.48	3.59	4.16	2.11	2.69	4.38	5.26
HIF1A	360	1789	1472	334	2007	1430	6.01	4.28	2.56	3.25	3.16	−0.08	2.01	1.52	2.08	1.59	4.32	3.12
JUNB	1028	215	691	953	241	671	0.25	0.70	3.01	2.33	2.83	1.16	−0.32	0.41	−1.48	−0.75	−3.07	−1.47
ATFA	1536	324	766	1424	364	744	0.26	0.52	3.17	2.50	2.88	1.60	0.10	0.57	−1.50	−1.03	−3.10	−2.01
GRO3	1596	304	447	1479	341	434	0.23	0.29	3.19	2.47	2.65	1.66	0.03	−0.21	−1.62	−1.87	−3.36	−3.66
SUPT3H	1266	208	439	1173	234	426	0.20	0.36	3.08	2.32	2.64	1.32	−0.35	−0.23	−1.68	−1.56	−3.48	−3.05
CCND1	1849	320	762	1713	360	740	0.21	0.43	3.26	2.50	2.88	1.85	0.11	0.58	−1.73	−1.27	−3.59	−2.48
Average	886	732	846	821	821	821	1.02	1.54	2.58	2.46	2.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00
standev	3574	3030	2429	3313	3400	2358	0.86	0.73	0.36	0.38	0.28	0.98	0.97	0.95	0.48	0.51	1.00	1.00
Norm factor	0.93	1.12	0.97															
Grand average	821																	

Comparison of two types of data normalization: global normalization and Z score transformation. Raw intensity data for three separate experimental conditions are shown under the heading Original data (Ct, control; PI, PMA + I; P28, PMA + 28). Global normalization of the corresponding raw intensity data is featured in the columns under the heading Normalized data. Changes in gene expression as calculated for globally normalized data are featured in the columns under the heading Fold change. Data in the columns under the heading Log data correspond to the $\log_{10}$ transformation of the original raw intensity data in preparation for Z score transformation, the results of which are reported in the columns under the heading: Z transformed data. Changes in gene expression between different Z transformed datasets are first calculated as simple differences between the corresponding Z scores (and reported in the columns under the heading Z diffs) and then divided by the standard deviation of each Z difference dataset and reported in the columns under the heading: Z ratios. The data shown here is a subset of the entire 1132 genes featured on the NIA-Immunoarray, selected for genes that exhibited significant increase or decrease across the experimental parameters. The data is sorted in descending order on the PI-Ct column under the Z ratios heading.

or condition 2. *P* values can be assigned to the calculated Z test value by consulting the critical Z value for a two-tailed test in a standard normal distribution table.

SAM

SAM (Significance Analysis of Microarrays; software from Stanford University labs[13]) analysis was performed on Z score data for two class-unpaired data using the default settings. The samples chosen for analysis came from Donor 1 and included three labeling replicates for control RNA, and two labeling replicates each for the PMA+I and PMA + 28 samples. The SAM procedure combines the calculation of a *t*-test statistic value for each gene with subsequent permutation analysis and the calculation of a false discovery rate (FDR). Significant gene changes were arbitrarily selected at SAM (d) score values greater than or equal to ± 1.46 (this value yielded the best balance between absolute number of significant calls and the lowest predicted false discovery rate (FDR) for the dataset tested).

Cluster Analysis

Hierarchical clustering of experimental variation in gene expression was determined using software programs developed at Stanford University.[16] The cluster algorithm was set to complete linkage clustering using the uncentered Pearson correlation.

Results

We have tested the use of the Z score transformation method for the normalization of microarray data across a wide range of hybridization results.[9,17–19] The sample dataset for this report is taken from a study of the gene expression differences (unpublished data) between two alternative pathways of T-cell stimulation: phorbol myristate plus ionomycin (PMA+I) or phorbol myristate plus anti-CD28 antibody (PMA + 28) following 2 hours of stimulation. The two conditions were compared to controls using either global normalization or Z score transformation and the results were used to compare several different downstream analytical techniques for estimating significant changes in gene expression. The complete dataset is available at http://www.grc.nia.nih.gov/branches/rrb/dna/dnapubs.htm.

Figure 1 shows a typical hybridization result from these experiments using NIA-Immunoarray filters. Dramatic increases in gene expression between control and stimulated T-cells are illustrated by, for example, the obvious

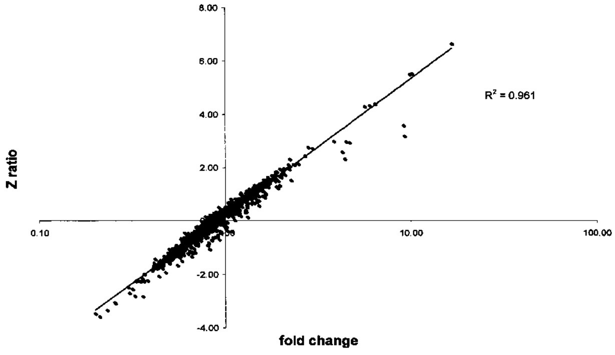


Figure 2. Scatter plot showing the linear relationship between measurements of changes in gene expression using either globally normalized data (fold change) or **Z** transformed data (**Z** ratio) on a lognormal scale. The r^2 value of the linear regression = 0.961.

Table 2. Estimates of Significant Changes in Gene Expression

	Global normalization		Z transformation		Genes in common†	
Treatment	PMA + 1*	PMA + 28†	PMA + 1*	PMA + 28†	PMA + 1*	PMA + 28†
Genes (no.) up-regulated	32	226	22	24	22	24
Genes (no.) down-regulated	67	19	20	25	20	20

*PMA + Ionomycin
†PMA + anti-CD28
The number of genes up-regulated and down-regulated by treatments shown according to traditional fold ratio or Z ratio. Significance levels for fold change estimates were set at two-fold changes up or down: for Z ratios the significance level was set at ± 2 Z ratio (approximate 95% confidence level).

increase in the hybridization signal for interferon-γ (IFN-gamma). While it is clear from visual inspection that both forms of T-cell activation result in an up-regulation of IFNg, it would also appear from the comparative signal strengths that PMA+I exerts a stronger effect on IFNg gene expression than does PMA + 28. However, it is not possible to reach this conclusion quantitatively without performing some form of normalization procedure before comparing the intensity values derived from the two hybridizations.

Table 1 demonstrates the process for deriving normalized data values using either mean-adjusted (global) normalization or Z score transformation. The data shown here are a subset of the total number of 1132 genes on the NIA Immunoarray, selected for genes that exhibited the most significant increases or decreases between experimental and control samples (the data are sorted in descending order on the PI-Ct column under the Z ratios heading). The first three columns following the gene names contain the raw intensity data before any adjustment and provide the basis for both types of normalization procedures.

Estimates of significant changes in gene expression using globally normalized data were calculated by traditional fold-ratio methods and are shown for each comparison (Table 1). Z ratio calculations were also determined for the same genes following Z score transformation. There is a close but not exact correspondence in rank order between the most highly up- or down-regulated genes calculated by either method (see,

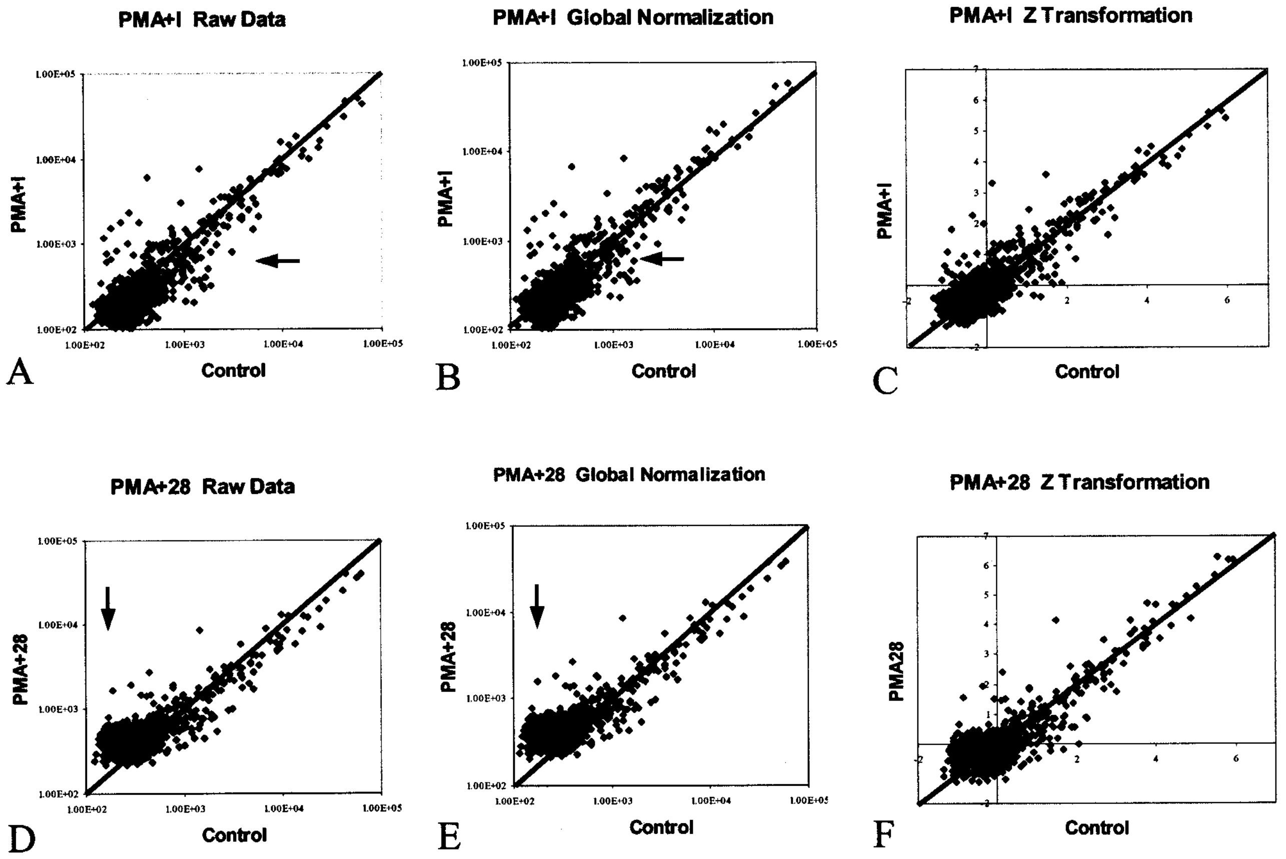


Figure 3. **A–C:** Scatter plots of treated PMA plus ionomycin, PMA+I; **D–F:** PMA plus anti-CD28, PMA + 28; untreated cells, control (all panels). **A** and **D** compare the raw data. **B** and **E** compare globally normalized data. **C** and **F** compare **Z** score data. The distortion in the data (**E**) resulted in 226 genes being called as up-regulated by global normalization in the PMA + 28 treatment. As shown for **Z** score transformation, this distortion is largely reduced (**F**) resulting in 24 genes being up regulated. **Arrows** indicate areas of data distortion introduced by the PMA + 28 hybridization filter.

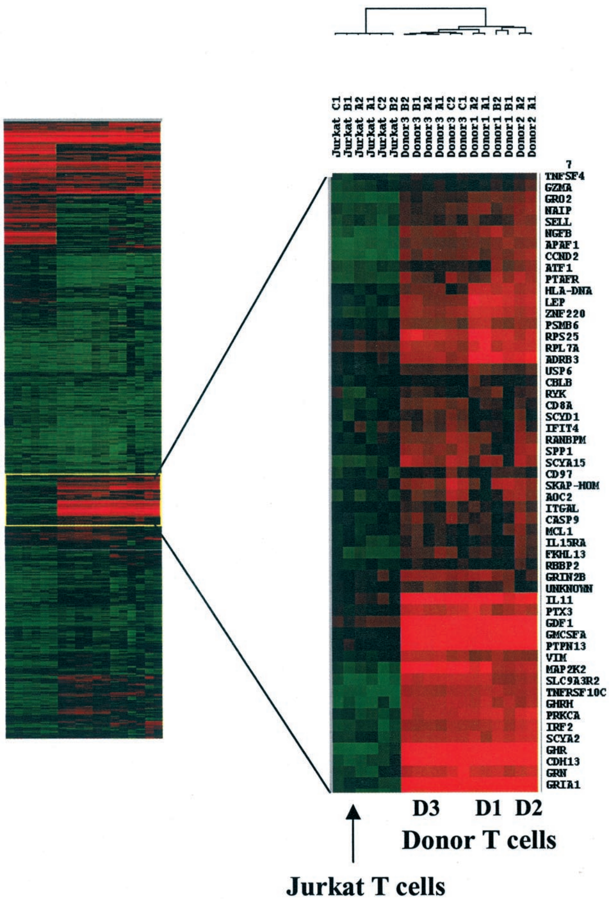


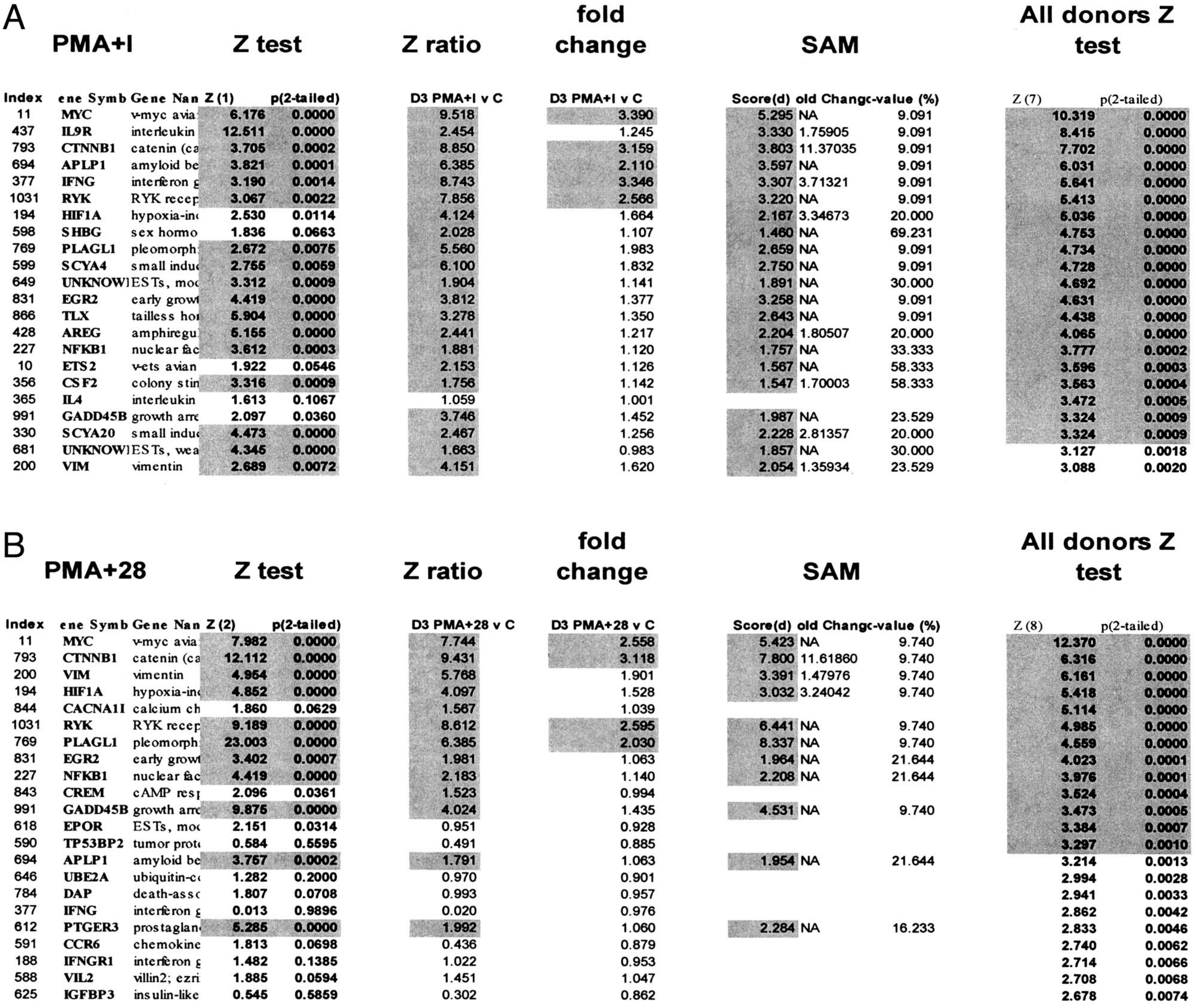
Figure 4. Cluster analysis of **Z** score data, sorting both genes and experiments simultaneously. Shown here is data from control cells only. Samples include microarray data from both Jurkat as well as donor PBTs (D1, D2, D3). RNA labeling replicates are alphabetical (**A**, **B**, **C**), duplicate data from the same filter are numeric(1,2). Scale of red to green equals higher to lower gene expression.

for example, VIM, P28 *vs* Ct). Figure 2 illustrates the relationship between the complete datasets of measured Z ratios and fold changes for this condition in the form of a lognormal scatter plot. Linear regression analysis shows that the data fit to a straight line with an r^2 value of 0.96. Thus, for this example, estimates of changes in gene expression by the two methods are virtually identical using two distinct normalization procedures (global or Z score transformation) as well as two distinct significance estimates (fold change or Z ratios). This equivalence suggests that the Z normalization as well as Z ratio calculations adequately mirror standard mean-adjusted normalization and subsequent fold change estimates of gene expression.

One advantage of Z score transformation appears to be the ability to minimize the distortion introduced to microarray datasets as a result of the occasional artifacts encountered during labeling, hybridization, and image acquisition. This point is illustrated by the comparison of PMA + 28 data results in Table 2 where 226 genes were designated as up-regulated using global normalization *versus* 24 genes that are reported as up-regulated by the Z ratio method. Figure 3 examines this effect by illustrating the results obtained using either global normalization or Z score transformation as applied to the two separate datasets: PMA+I versus control (Figure 3, A–C) or PMA + 28 versus control (Figure 3, D and E). The top panels show the original and normalized data for PMA+I. In this case, both normalization methods appear to balance the data symmetrically and, therefore, it is not surprising that the resulting gene changes are found to be highly similar, as previously mentioned (Figure 2). In contrast, the original and normalized data for PMA + 28 (Figure 2, D–F), demonstrate that Z score transformation but not global normalization is capable of correcting the data distortion (indicated by arrows) introduced by the PMA + 28 hybridization data. Although the exact cause for this data distortion is unclear (but neither is it uncommon; a false positive rate associated with low gene expression has been noted by other investigators[20]), it clearly accounts for the skewed results reported for PMA+I up-regulation in Table 2, mostly as the result of a bias in the data in the PMA + 28 direction at lower expression levels. Furthermore, it should also be clear that inclusion of the PMA + 28 data in any extended set of experimental comparisons would continue to exert a distorting effect. In general, Z normalization shows greater stability as a result of examining where each gene intensity falls in the overall distribution of values within a given array as opposed to adjusting all of the genes in an array by a single common value.

Z scores can be used directly in any number of analysis formats including cluster analysis as demonstrated in Figure 4. Z scores allow clustering and other analytical methods to be used on corrected values that are directly related to original intensity values as well as on comparative ratios. The Z scores used here were generated from a wider range of experiments than from the study of T cell activation previously shown and include data from three independent human T-cell donors as well as from cultured Jurkat T cells. The unsupervised clustering algorithm (both genes and experiments were allowed to cluster independently) clearly separates not only the two distinct cell types (Jurkat and donor T cells) but also precisely sorts the primary T-cell cultures by individual donor. The high correlation of clustered experimental groups to the original experimental identities suggests that Z score transformation by itself is a reliable normalization process, providing a basic unit of comparison between a wide range of samples being analyzed (within a single microarray type).

In addition to the Z ratio method for computing significant changes in gene expression using Z scores, we have also tested several additional statistical techniques including the Z test[12] and SAM analysis[13] and compared the results together as illustrated in Figure 5. The data shown here is a sampling of the most significant genes up-regulated by treatment with either PMA+I (A) or PMA + 28 (B). The majority of all significant changes in gene expression as calculated by any of the tested methods can be seen in Figure 5. Gene changes exceeding specified significance thresholds are shaded in gray. The Z test, Z ratio, fold change, and SAM analysis data in this example were taken from the same set of experiments using T cells harvested from a single donor. These results



A

PMA+I			Z test		Z ratio	fold change	SAM			All donors Z test	
Index	ene Symb	Gene Nan	Z (1)	p(2-tailed)	D3 PMA+I v C	D3 PMA+I v C	Score(d)	old Changc	-value (%)	Z (7)	p(2-tailed)
11	MYC	v-myc avia	6.176	0.0000	9.518	3.390	5.295	NA	9.091	10.319	0.0000
437	IL9R	interleukin	12.511	0.0000	2.454	1.245	3.330	1.75905	9.091	8.415	0.0000
793	CTNNB1	catenin (ca	3.705	0.0002	8.850	3.159	3.803	11.37035	9.091	7.702	0.0000
694	APLP1	amyloid be	3.821	0.0001	6.385	2.110	3.597	NA	9.091	6.031	0.0000
377	IFNG	interferon g	3.190	0.0014	8.743	3.346	3.307	3.71321	9.091	5.641	0.0000
1031	RYK	RYK recep	3.067	0.0022	7.856	2.566	3.220	NA	9.091	5.413	0.0000
194	HIF1A	hypoxia-in	2.530	0.0114	4.124	1.664	2.167	3.34673	20.000	5.036	0.0000
598	SHBG	sex hormo	1.836	0.0663	2.028	1.107	1.460	NA	69.231	4.753	0.0000
769	PLAGL1	pleomorph	2.672	0.0075	5.560	1.983	2.659	NA	9.091	4.734	0.0000
599	SCYA4	small indu	2.755	0.0059	6.100	1.832	2.750	NA	9.091	4.728	0.0000
649	UNKNOWI	ESTs, moc	3.312	0.0009	1.904	1.141	1.891	NA	30.000	4.692	0.0000
831	EGR2	early growt	4.419	0.0000	3.812	1.377	3.258	NA	9.091	4.631	0.0000
866	TLX	tailless hoi	5.904	0.0000	3.278	1.350	2.643	NA	9.091	4.438	0.0000
428	AREG	amphiregu	5.155	0.0000	2.441	1.217	2.204	1.80507	20.000	4.065	0.0000
227	NFKB1	nuclear fac	3.612	0.0003	1.881	1.120	1.757	NA	33.333	3.777	0.0002
10	ETS2	v-ets avian	1.922	0.0546	2.153	1.126	1.567	NA	58.333	3.596	0.0003
356	CSF2	colony stin	3.316	0.0009	1.756	1.142	1.547	1.70003	58.333	3.563	0.0004
365	IL4	interleukin	1.613	0.1067	1.059	1.001				3.472	0.0005
991	GADD45B	growth arre	2.097	0.0360	3.746	1.452	1.987	NA	23.529	3.324	0.0009
330	SCYA20	small indu	4.473	0.0000	2.467	1.256	2.228	2.81357	20.000	3.324	0.0009
681	UNKNOWI	ESTs, wea	4.345	0.0000	1.663	0.983	1.857	NA	30.000	3.127	0.0018
200	VIM	vimentin	2.689	0.0072	4.151	1.620	2.054	1.35934	23.529	3.088	0.0020

B

PMA+28			Z test		Z ratio	fold change	SAM			All donors Z test	
Index	ene Symb	Gene Nan	Z (2)	p(2-tailed)	D3 PMA+28 v C	D3 PMA+28 v C	Score(d)	old Changc	-value (%)	Z (8)	p(2-tailed)
11	MYC	v-myc avia	7.982	0.0000	7.744	2.558	5.423	NA	9.740	12.370	0.0000
793	CTNNB1	catenin (ca	12.112	0.0000	9.431	3.118	7.800	11.61860	9.740	6.316	0.0000
200	VIM	vimentin	4.954	0.0000	5.768	1.901	3.391	1.47976	9.740	6.161	0.0000
194	HIF1A	hypoxia-in	4.852	0.0000	4.097	1.528	3.032	3.24042	9.740	5.418	0.0000
844	CACNA1I	calcium ch	1.860	0.0629	1.567	1.039				5.114	0.0000
1031	RYK	RYK recep	9.189	0.0000	8.612	2.595	6.441	NA	9.740	4.985	0.0000
769	PLAGL1	pleomorph	23.003	0.0000	6.385	2.030	8.337	NA	9.740	4.559	0.0000
831	EGR2	early growt	3.402	0.0007	1.981	1.063	1.964	NA	21.644	4.023	0.0001
227	NFKB1	nuclear fac	4.419	0.0000	2.183	1.140	2.208	NA	21.644	3.976	0.0001
843	CREM	cAMP resp	2.096	0.0361	1.523	0.994				3.524	0.0004
991	GADD45B	growth arre	9.875	0.0000	4.024	1.435	4.531	NA	9.740	3.473	0.0005
618	EPOR	ESTs, moc	2.151	0.0314	0.951	0.928				3.384	0.0007
590	TP53BP2	tumor prot	0.584	0.5595	0.491	0.885				3.297	0.0010
694	APLP1	amyloid be	3.757	0.0002	1.791	1.063	1.954	NA	21.644	3.214	0.0013
646	UBE2A	ubiquitin-c	1.282	0.2000	0.970	0.901				2.994	0.0028
784	DAP	death-assc	1.807	0.0708	0.993	0.957				2.941	0.0033
377	IFNG	interferon g	0.013	0.9896	0.020	0.976				2.862	0.0042
612	PTGER3	prostaglan	5.285	0.0000	1.992	1.060	2.284	NA	16.233	2.833	0.0046
591	CCR6	chemokine	1.813	0.0698	0.436	0.879				2.740	0.0062
188	IFNGR1	interferon g	1.482	0.1385	1.022	0.953				2.714	0.0066
588	VIL2	villin2; ezri	1.885	0.0594	1.451	1.047				2.708	0.0068
625	IGFBP3	insulin-like	0.545	0.5859	0.302	0.862				2.678	0.0074

Figure 5. Comparison of several different methods for estimating significant changes in gene expression (highlighted by shaded values) using the same dataset. Data shown is a sampling of the most significant genes up-regulated by treatment with either PMA plus ionomycin (**A**) or PMA plus an antibody to the CD28 receptor (**B**). Significance thresholds for fold change were set at ≥ 2, for **Z** ratios at ≥ 1.5, for **Z** test at $P \leq 0.01$, for SAM at d > 1.45. SAM, fold changes, **Z** ratios, and **Z** test data are from a single donor. Labeling replicates (3 for the control sample RNA, 2 each for PMA+I and PMA + 28 sample RNAs) were first averaged for **Z** ratio and fold change estimates. These results are compared to a composite of replicated data (using averages of multiple individual donor RNA labelings) from three individuals (all donors **Z** test). The data are sorted in order of decreasing **Z** test value for the all donors column.

were then compared to a composite of replicated data from three individual donors (all donors Z test) by which the data are sorted in descending order.

An inspection of the results in Figure 5 shows that the fold changes calculated from globally normalized data appear to greatly underestimate the number of significant gene changes as predicted by all other methods. The Z test and Z ratio results appear to agree quite well with each other with some exceptions. The Z test was somewhat more conservative than the Z ratios for determining significant changes in gene expression. Both Z ratios and Z tests overlap with the significance determinations made by the SAM program. In addition, and perhaps, most importantly, the Z ratio, Z test and SAM predictions made from data from a single donor appear to reliably mirror the significance estimates made by performing the Z test on the combined experimental data derived from three individual donors. The significant overlap between three independent statistical analysis methods and the agreement between single and multiple donor data suggests that, in the aggregate, reliable changes in gene expression can be consistently detected by all three methods.

Discussion

We have addressed three issues of importance in microarray comparisons: normalization of microarray results, calculation of an appropriate gene comparison statistic, and the development of a flexible and archivable data storage method by application of the Z score transformation method specifically to microarray analysis. Z scores provide a useful measurement of gene expression that can be used in downstream analysis as proportional to the hybridization intensities from which they were derived. Z scores have been successfully used directly in hierarchical clustering, k-means clustering, self-organizing maps (SOM), principal component analysis (PCA), multidimensional scaling as well as in visualization programs such as GeneSpring (data not shown). Z scores (as well as the Z ratios and Z test statistics) can be rapidly calculated from raw data through the use of a Microsoft Excel spreadsheet available at http://www.grc.nia.nih.gov/branches/rrb/dna/dnapubs.htm.

Z scores provide a relative, semiquantitative estimate of gene expression levels and, as such, form the basis of comparison of hybridization intensity data among many experiments within the same array type. Direct inspection of Z score values in visualization analyses such as hierarchical clustering is aided by the fact that Z scores are proportional to the intensity of the original hybridization signal. The value of the Z score is directly reflective of the underlying differential hybridization values (ie, higher positive Z scores represent the most highly expressed genes, lower negative Z scores represent the least expressed genes). Thus Z scores provide a useful and intuitive method for visualizing and interpreting very large amounts of data in their natural biological context. This is in contrast to normalization strategies that express hybridization intensities as ratios of one sample to another (either experimental or to a common reference sample). The values derived by ratio normalization techniques are more difficult to interpret because they are always dependent on the normalizing sample from which they were derived. Positive and negative values in these analyses simply indicate their relationship to the normalizing sample rather than reflecting actual gene expression levels. Ratio normalization thus makes it difficult to compare many different experiments directly even when using the same array type.

Z ratios provide a relative measure of significant gene expression changes in pair-wise group comparisons. In this regard, Z ratios are the conceptual equivalent to Cy3/Cy5 ratios generated using two-color fluorescent techniques. Just as Z scores are used to analyze many different experiments in terms of relative intensity measurements, Z ratios can be used to compare significant changes in gene expression across a similarly wide range of experiments. The advantages of Z ratios are that they are directly comparable among many different experiments, rapidly calculated, and show good agreement (Figure 5) with more complex statistical analyses (eg, SAM analysis).

The application of Z test statistics to Z score microarray data was used to address additional requirements for a more rigorous statistical analysis than provided for solely by Z ratios. These improvements include, in the Z test, a SE method for balancing the effects of repeated measurement variation versus the statistical power afforded by replicate numbers. Because the Z test places a high value on low variability between experimental replicates, it tends to be more conservative than Z ratios for finding significant gene changes. Indeed, the impact on significance calculations of sample variation when using the Z test is such that it has mitigated the need, in our hands, for *a priori* outlier removal. The use of the Z test is facilitated directly in the Excel program, which provides one-tailed *P* values for the Z distribution (function = NORMSDIST).

The comparability of data are critical in the handling of the ever-increasing data streams being generated during ongoing microarray gene expression studies. Z scores will not, theoretically, be comparable outside of the array type from which they are generated since their value is specifically linked to a fixed population of genes. This limitation, however, is almost universal in the field of microarray techniques making cross-array and cross-platform comparisons difficult and inhibiting the growth of universal databases. A reliable method such as Z score transformation is, nonetheless, vital for intra-array comparisons in large studies using a stable focused array format (eg, NIA-Immunoarray).

Acknowledgments

We thank James Nagel and Dennis Taub for collaboration and development of the NIA Immunoarray, Paritosh Ghosh and Meredith Buchholz for preparation of the T cells and total RNA used to generate the sample datasets, and William Wood III for filter preparation.

References

1. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: Expression profiling using cDNA microarrays. Nat Genet 1999, 21:10–14
2. Brazma A, Vilo J: Gene expression data analysis. FEBS Lett 2000, 480:17–24
3. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H: Normalization strategies for cDNA microarrays. Nucleic Acids Res 2000, 28:E47
4. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002, 30:E15
5. Ishii K, Sasaki M, Matsui M, Sakamoto S, Yamaji S, Hayashi N, Mori T, Kitagaki H, Hirono N, Mori E: A diagnostic method for suspected Alzheimer's disease using H(2)15O positron emission tomography perfusion Z score. Neuroradiology 2000, 42:787–794
6. Guilford JP: Structure-of-intellect abilities in preliterate children and the mentally retarded. Monogr Am Assoc Ment Defic 1973, 1:46–58
7. Fruchter B, Fruchter DA: Factor content of the WAIS with separate digits-forward and digits-backward scores for a borderline and mentally retarded sample. Monogr Am Assoc Ment Defic 1973, 1:67–73
8. Becker KG, Barrett T, Vawter MP, Wood WH, Cheadle C: cDNA arrays in neuroscience: membrane based assays. Society for Neuroscience, New Orleans, LA, 2000 (Program 232.1)
9. Vawter MP, Barrett T, Cheadle C, Sokolov BP, Wood WH, Donovan D, Webster M, Freed WJ, Becker KG: Application of cDNA microarrays to examine gene expression differences in schizophrenia. Brain Res Bull 2001, 55:641–650
10. Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de la Chapelle AA, Krahe R: Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. Proc Natl Acad Sci USA 2001, 98:1124–1129
11. Lemkin PF, Thornwall GC, Walton KD, Hennighausen L: The microarray explorer tool for data mining of cDNA microarrays: application for the mammary gland. Nucleic Acids Res 2000, 28:4452–4459
12. Nadon R, Woody E, Shi P, Rghei N, Hubschle H, Susko E, Ramm P (2002). Statistical inference in array genomics. Microarrays for the Neurosciences. Edited by Geschwind D, Gregg J. Cambridge, MIT Press, pp 109–140

13. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001, 98:5116–5121
14. Chomczynski P, Sacchi N: Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 1987, 162:156–159
15. Barrett T, Cheadle C, Wood WHI, Donovan D, Freed WJ, Becker KG, Vawter MP: Assembly and use of a broadly applicable neural cDNA microarray. Restor Neurol Neurosci 2001, 18:127–135
16. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998, 95:14863–14868
17. Cho YS, Meyoung-Kim M, Cheadle C, Neary C, Becker KG, Cho-Chung YS: Antisense DNAs as multisite genomic modulators identified by DNA microarray. Proc Natl Acad Sci USA 2001, 98:9819–9823
18. Mayne M, Cheadle C, Soldan SS, Cermelli CC, Yamano Y, Akhyani N, Nagel JE, Taub DD, Becker KG, Jacobson S: Gene expression profile of herpesvirus-infected T cells using immuno-microarrays: up-regulation of pro-inflammatory genes. J Virol 2001, 75:11641–11650
19. Truckenmiller ME, Vawter M, Cheadle C, Coggiano M, Donovan DM, Freed WJ, Becker KG: Gene expression profile in early stage of retinoic acid-induced differentiation of human SH-SY5Y neuroblastoma cells. Restorative Neurology and Neuroscience 2001, 18:67–80
20. Baldi P, Long AD: A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene change. Bioinformatics 2001, 17:509–519 .0