Journal of Cellular Biochemistry Supplement 37:120–125 (2001)

Feature Extraction and Normalization Algorithms for High-Density Oligonucleotide Gene Expression Array Data

Eric E. Schadt,[1,4]* Cheng Li,[3] Byron Ellis,[2] and Wing H. Wong[2,3]**

[1]Department of Biomathematics, University of California, Los Angeles, California
[2]Department of Statistics, Harvard University, Cambridge, Massachusetts
[3]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts
[4]Departments of Informatics, Rosetta Inpharmatics, Kirkland, Washington

Abstract Algorithms for performing feature extraction and normalization on high-density oligonucleotide gene expression arrays, have not been fully explored, and the impact these algorithms have on the downstream analysis is not well understood. Advances in such low-level analysis methods are essential to increase the sensitivity and specificity of detecting whether genes are present and/or differentially expressed. We have developed and implemented a number of algorithms for the analysis of expression array data in a software application, the DNA-Chip Analyzer (dChip). In this report, we describe the algorithms for feature extraction and normalization, and present validation data and comparison results with some of the algorithms currently in use. J. Cell. Biochem. Suppl. 37: 120–125, 2001.

Key words: feature extraction; normalization; oligonucleotide array; gene expression

Monitoring gene expression using high-density microarrays is a key technology in the study of cell functions and the associated biochemical pathways, candidate gene identification, cellular response to drug compounds, and classification of disease states [Wodicka, 1997; Eisen, 1998; Zhu, 1998; Alon, 1999; Golub, 1999; Tamayo, 1999]. Published methods have largely focused on enhancing the technology itself and the corresponding experimental protocols [Schena, 1995; Lockhart, 1996; Shalon, 1996; Mahadevappa and Wodicka, 1999], and on developing higher-level analysis methods such as clustering and classification. Chen (1997) detailed algorithms for image segmentation and confidence intervals for expression ratios for cDNA microarray data. For oligonucleotide array data, a thorough investigation of such low-level analysis is lacking.

We present novel algorithms for two critical steps in the analysis of oligonucleotide expression arrays such as the Affymetrix GeneChip® probe arrays [Lockhart, 1996; Lipshutz et al., 1999]. Specifically, we describe our methods for segmenting array images and computing feature intensities, and for normalizing two or more arrays. It is well known that the comparison of gene expression results across experiments relies crucially on having an effective normalization scheme. Validation data for each of the algorithms are presented and the algorithms are compared against alternative algorithms currently in use. It is found that the new algorithms offer substantial gains in reducing replication variability and in enhancing estimation of expression ratios. Finally, we describe our software, the DNA-Chip Analyzer, for the analysis of oligonucleotide expression array data.

Grant sponsor: NSF; Grant numbers: DMS-9703918, DBI-9904701.

*Correspondence to: Eric E. Schadt, Department of Biomathematics, University of California, Los Angeles, CA and Departments of Informatics, Rosetta Inpharmatics, Kirkland, WA. E-mail: eschadt@rosetta.org

**Correspondence to: Wing H. Wong, Department of Biostatistics, Harvard School of Public Health, 655 Huntington, Ave, Boston, MA 02115.
E-mail: wwong@stat.harvard.edu

Received 21 August 2001; Accepted 22 August 2001

DOI 10.1002/jcb.10073

RESULTS

Feature Extraction

The feature extraction process involves defining a grid that identifies each of the features on an array, and then computing an intensity value for each feature. The oligonucleotide array image contains easily recognizable alignment features at each corner (Fig. 1/Image 1). Since the physical dimension of each feature is known, once the positions of the alignment features are determined, a simple bilinear transformation is used to map the positions of each feature in the array. This defines the basic grid that segments the raw image into individual features. To compute the intensity value for a feature, current methods [Lockhart, 1996] take the 75th percentile (TP75) of the pixel intensities for the feature after removing the boundary pixels. In Figure 1 (Image 2), the intensity of a perfect match (PM) feature is seen to be much higher than the corresponding mismatch (MM) feature. Because of blurring, the pixels near the PM/MM border are distorted in their value, resulting in an upward bias when the TP75 algorithm is used to compute the MM intensity. Another example given in Figure 1 (Image 3) shows how a misalignment of the basic grid results in a failure to extract the central part of the true feature. To address these problems, we have implemented an adaptive pixel selection algorithm (APS). The first step is masking pixels with extreme intensities (i.e., removal of pixels more than 3 standard deviations from the mean pixel value within a feature). Then, the edge whose removal results in the greatest reduction in the coefficient of variation (CV) of the remaining pixels, is removed, if the reduction is judged to be statistically significant. This is repeated until no further significant reduction in the CV can be achieved or until the feature size has been reduced to a predefined minimum (by default, 4×4 pixels). In addition, we constrain the pixel selection process by forcing adjacent subregions selected by this process (corresponding to adjacent features) to be separated by at least two pixels. This procedure tends to select the most homogenous group of pixels whose mean value is used to represent the intensity for the given feature.

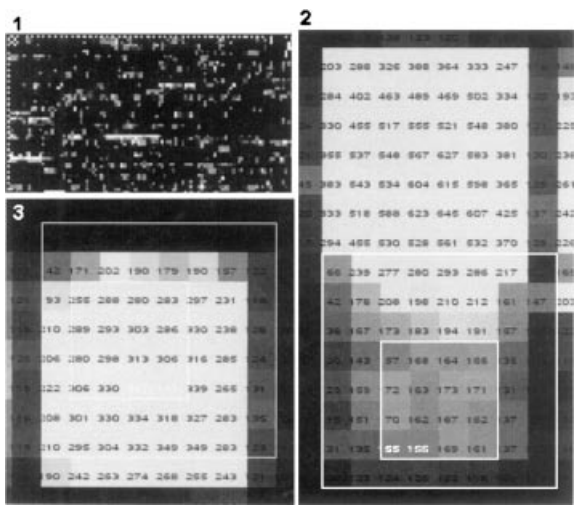

Fig. 1. Image 1 represents a feature-level view, generated by dChip, of the upper left corner of a high-density Affymetrix Hu6800 probe array. The checker-board pattern generated by the alignment features is easily detected. Image 2 depicts a PM (top of image) and MM (bottom of image) feature where the bright intensities are seen to blur the PM/MM boundary, resulting in an upward bias when the TP75 algorithm is used to compute the MM intensity (198 vs. the APS computed value of 154). The outer box in image 3 is part of the basic grid that defines the boundary for each feature. It is clear the basic grid is not properly centered over this feature. The inner box of image 3 is the region selected by the APS algorithm.

The APS algorithm was compared to the TP75 algorithm by examining twelve replicate oligonucleotide arrays. These replicate data were generated by hybridizing the same cRNA hybridization cocktail onto six high-density Affymetrix Hu6800 probe arrays and six "A" probe arrays from the low-density Affymetrix Hu6800 four-chip set. The twelve arrays were normalized using the IDS/GCVSS normalization algorithm described below. It is reasonable to expect, after normalization, the intensities for any given feature across the twelve replicates to be roughly equal, since the same sample was hybridized onto each array. For each feature across the twelve replicate arrays, we computed the feature-intensity standard deviations (SD) after using the TP75 and the APS algorithms to compute the feature intensities. A good feature extraction algorithm should lead to a small SD among the replicates. Analyzing the twelve replicate probe arrays, we found that 63% of the APS feature-intensity SDs were significantly smaller than the corresponding TP75 feature-intensity SDs, and that in these cases, the mean TP75-computed SD/APS-computed SD ratio was 1.53. On the other hand, the mean ratio between the APS-computed SDs and the TP75-computed SDs was only 1.31, when the APS-computed SD was larger than the TP75-computed SD. Thus the APS algorithm leads to a 17% reduction in the intensity standard deviation across replicate

arrays, when compared to the TP75 algorithm. Given that these feature intensity calculations form the basis for all future analyses, any method that significantly reduces the measurement error will serve to increase the sensitivity and specificity of these types of experiments.

Normalization

Normalizing multiple probe arrays to allow direct array-to-array comparisons presents one of the greatest challenges in expression array data analysis. Current methods include (1) linear normalization and its extension by nonlinear regression and (2) methods based on housekeeping genes or staggered spike-in controls. Linear normalization, the most popular normalization method currently in use, assumes the intensities between two or more arrays are related as a straight line with a zero y-intercept. It leads to multiplication by a scaling factor (slope of the line) to make the mean of the experiment chip, the same as that of the baseline chip. Although simple and robust, this method has the drawback that it cannot adjust for nonlinear relations. Figure 2 illustrates a situation where the slope in the low intensity region (of the scatter plot of PM/MM differences between two arrays) is substantially different from the slope in the high intensity region. In examining many arrays, we have found that a 10–50% difference in slope values is quite common. A natural modification of the linear method is to fit a nonlinear regression of the baseline array values on the experiment array values (Fig. 2). An implementation of such a procedure using smoothing splines with generalized cross-validation (GCVSS) [Wahba, 1990] was described in Schadt [1999]. We will see, however, that such a procedure is inadequate if the expression profiles of the two arrays are very different.

It has also been suggested [Ermolaeva, 1998] that normalization between arrays can be based on a set of "housekeeping" genes. However, in profiling human and murine tissue samples, we have found many of the genes currently used as housekeeping genes (e.g., β actin, glyceraldehyde-3-phosphate dehydrogenase, transferrin receptor, signal transducer, and activator of transcription 1, among others) to have ranges of differential expression similar to other genes whose differential expression patterns are deemed biologically relevant to the system under study. We have also investigated establishing normalization relationships using control cRNAs for bacterial and phage genes (e.g., BioB, BioC, BioD, and cre), which are consistently added to hybridization mixtures at known concentrations. However, these controls are often prepared in bulk and completely independently of the sample being profiled, and so, the normalization relation between the controls on different arrays typically does not

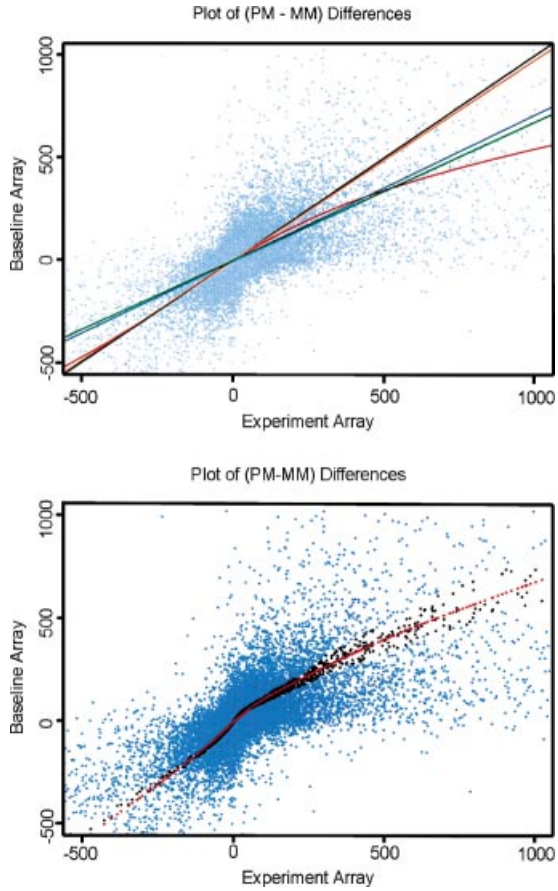

Fig. 2. **Panel 1** illustrates the distribution differences that can exist between the low and high feature intensities. The PM/MM differences from two murine Affymetrix Mu6500SubA probe array experiments are plotted. The green line in Panel 1 is the line generated by the LR normalization method, the red curve is generated by the GCVSS method, and the orange and blue lines are generated by applying the LR method to low and high differences, respectively (the low/high cutoff was determined empirically from the GCVSS curve, and was taken to be approximately 20). The slope of the green line is 1.47, while the slope of the orange line is only 1.04 (a 30% reduction). The GCVSS line matches the orange line at the low end and the blue and green lines at the high end, although it is clear the data between the two experiments are not really linearly related (the R^2 for the green line is only 53%; after normalization using the IDS/GCVSS technique the R^2 jumped to 78%). **Panel 2** shows the same data presented in Panel 1. The black points represent the differences chosen by the IDS technique and the red line is the GCVSS fit of the invariant differences between the baseline and experiment arrays.

reflect the true normalization relation for the biologically relevant genes of interest.

To address these problems, we have developed an invariant difference selection algorithm (IDS) that chooses a subset of PM/MM intensity differences to serve as the basis for fitting a normalization relation. A set of probes are said to be invariant if the ordering of these probes according to the PM/MM differences in the experiment array, is the same as that in the baseline array. Intuitively, if a gene is truly differentially expressed, then the PM/MM differences for this gene are more likely to have different ranks relative to the other probes, and hence they are not likely to be included in a large invariant set. Although the maximal invariant set can be computed using a dynamic programming algorithm (not presented), the resulting set is typically too small to form a reliable normalization curve. Our IDS algorithm finds an approximately invariant set of differences that is not necessary maximal, but is large enough for reliable fitting of the normalization relation. The IDS algorithm uses the following expressions to determine the approximately invariant set:

$$R_i = \frac{[L(B_i + E_i) + H(2N - B_i - E_i)]}{2N}$$

$$D_i = \frac{2|B_i - E_i|}{(B_i + E_i)}$$

where L and H are the rank difference thresholds for the low and high ends of the difference intensity range, B_i and E_i are the ranks for the ith difference of the baseline and experiment arrays, and N is the total number of differences that were ordered in the current iteration of the algorithm. R_i defines the threshold for difference intensity i by linearly interpolating the threshold between a low difference intensity threshold, given by L, and a high difference intensity threshold, given by H. This interpolation is needed because there are many more points at low intensities than at high intensities, so we can enforce a more strict threshold policy at the low end, but must relax this constraint at the high end to ensure enough points are obtained in this range to reliably establish the normalization relation. D_i is the rank difference test statistic used to determine if the ith difference should be included in the invariant set, for the current iteration of the algorithm. The ith difference is considered approximately invariant if $D_i < R_i$ This selection process then repeats, taking the current set of approximately invariant differences as input, until all differences meet the threshold criteria. Once the approximately invariant set of differences has been selected, the normalization curve is constructed by applying the GCVSS technique to the invariant set (Fig. 2).

In comparing our IDS/GCVSS normalization technique to the linear regression (LR) and GCVSS methods, we established that one normalization method worked better than another if (1) the method minimized the PM/MM intensity difference variances across a series of replicate arrays and (2) the method preserved expression ratios in simulated data. The first criterion ensures that genes known to have identical expression levels will have observed levels as close to identical as possible. The second criterion ensures that criterion 1 is not achieved at the expense of destroying the very biological variation the technology aims to detect. Table I presents the results of

TABLE I. Results of Comparing Average Difference Intensities Across a Series of 12 Replicate Probe Arrays Using No Normalization (UN), Linear Normalization (LR), Smoothing Spline Normalization (GCVSS), and the Invariant Difference Selection/Smoothing Spline (IDS) Methods

> Relation	UN	LR	GCVSS	IDS
UN	N/A	(0.96, 1.2)	(0.85, 1.7)	(0.74, 1.8)
LR	(0.04, 1.0)	N/A	(0.76, 1.4)	(0.65, 1.5)
GCVSS	(0.15, 1.3)	(0.24, 1.2)	N/A	(0.46, 1.4)
IDS	(0.26, 1.2)	(0.35, 1.3)	(0.54, 1.5)	N/A

The first number in each value pair of the table represents the percentage of standard deviations, computed across the 12 replicates for each of the genes that were larger when the normalization technique listed in the leftmost column was compared against the normalization technique listed along the top row. For instance, in cell (LR, IDS) the value pair (0.65, 1.5) indicates that 65% of the difference standard deviations were larger when LR normalization was used compared to the difference standard deviations when the IDS method was used, and of those that were larger, the median LR/IDS ratio of standard deviations was 1.5, i.e., for 65% of the genes, half of those had an LR standard deviation that was more than 1.5 times larger than the corresponding IDS standard deviations

TABLE II. Results From Two Sets of Simulated Expression-Ratio Data

> Relation	LR.1/LR.2	GCVSS.1/GCVSS.2	IDS.1/IDS.2
LR.1/LR.2	N/A	(0.08, 1.2)/(0.15, 1.3)	(0.94, 9.3)/(0.61, 2.1)
GCVSS.1/GCVSS.2	(0.92, 3.2)/(0.85, 3.2)	N/A	(0.99, 16.5)/(0.95, 3.9)
IDS.1/IDS.2	(0.06, 1.4)/(0.39, 1.4)	(0.01, 1.1)/0.05, 1.3)	N/A

In the first set, 300 genes that were consistently detected as present across the 6 low-density replicate probe arrays and 600 from the high-density replicate probe arrays were randomly selected; 6 sets containing 50 genes each for the low-density arrays and 100 genes each for the high-density arrays were then generated by randomly selecting, without replacement, from the sets of 300 and 600 randomly selected genes. The PM/MM differences comprising each of the genes in each of the sets were then multiplied by 2.0, 0.5, 4.0, 0.25, 6.0, and 0.17, respectively, to simulate fold changes between samples. The 12 original replicate probe arrays as well as the 12 modified replicate probe arrays were then normalized using the normalization techniques listed in Table I. The same procedure was applied to the second set, except that 320 genes from the low density arrays and 640 genes from the high-density arrays were randomly selected, and then 16 sets of 20 genes/40 genes each were formed and the corresponding difference intensities were multiplied by 2.0, 0.5, 2.5, 0.4, 3.0, 0.33, 3.75, 0.27, 4.0, 0.25, 5.0, 0.20, 6.0, 0.17, 7.0, and 0.14, respectively, yielding a very diverse differential expression pattern in which as many as 18% of the genes on the arrays were forced to be differentially expressed. The standard deviations of the differences between the true fold change and the observed fold change after normalization were computed for each of the modified genes. The first value in each value pair represents the percentage of standard deviations that were larger when the normalization technique listed in the leftmost column was compared against the normalization technique listed along the top row. For instance, the first value pair (0.94, 9.3) in cell (LR.1/LR.2, IDS.1/IDS.2) indicates that for the first set of data, 94% of the LR.1 computed standard deviations were larger than the IDS.1 computed standard deviations, and that of those that were larger, the median LR.1/IDS.1 ratio of standard deviations was 9.3, i.e., half of the time, the LR. 1 standard deviation is more than 9.3 times larger than the IDS.1 standard deviation.

validating the IDS/GCVSS method against the GCVSS method and the popular LR method. The validation was carried out on the same set of 12 replicate probe arrays discussed above. These results indicate that the GCVSS and IDS/GCVSS approaches are reasonably similar and do a better job making the average of the PM/MM differences for a particular gene, i.e., the average difference intensities, across these replicate arrays more consistent than doing nothing at all or than using the LR method. However, it becomes clear in Table II that the GCVSS technique makes the average difference intensities more consistent by destroying the very biological variation the technology aims to detect, i.e., the GCVSS technique is too sensitive to the relatively small number of genes that change. On the other hand, the

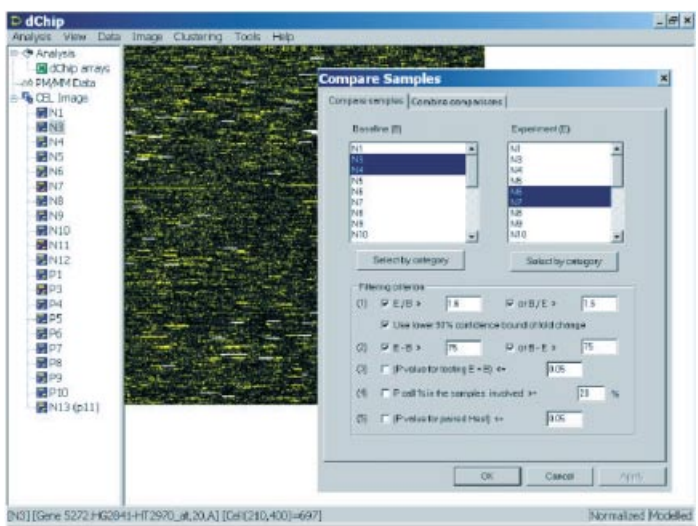

Fig. 3. Snapshot of the dChip software illustrating the user interface. The left pane demonstrates the use of a Microsoft Explorer-like tree control to organize arrays. Once the arrays have been processed and grouped, higher-level statistical tests on the data can be performed, such as hierarchical clustering and sample comparisons. Data used in the figure is courtesy of Andrea Richardson and Dirk Iglehart.

IDS/GCVSS technique continues to perform better than the other techniques, by not only minimizing the average difference intensity variation across replicates, but also minimizing the deviation from the true fold change values in the simulated data.

DISCUSSION

We have presented algorithms for feature extraction and normalization and demonstrated that these algorithms lead to improved computation of feature intensities and expression ratios. Because many important decisions on whether a gene should be pursued as a candidate for a particular biological system under study are directly based on the expression ratios as well as on the differential expression calls made by software such as the Affymetrix GeneChip software, algorithms that provide for more accurate estimates of these derived statistics will be of great value to users of this technology.

We have implemented the algorithms described in this article in our software application, the DNA-Chip Analyzer (dChip). In addition to the implementations of the algorithms described here, dChip also performs image gradient and artifacts correction, model-based expression index calculation, array and probe outlier detection [Wong and Li 2001]. dChip has a user interface that allows arrays and experiments to be logically grouped, and provides higher level group comparison functions and hierarchical clustering (Fig. 3). The software is available at www.dchip.org

REFERENCES

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750.

Chen Y, Dougherty ER, Bittner ML. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J Biomed Opt 2:364–374.

Eisen MB, Spellman P, Brown P, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868.

Ermolaeva O, Rastogi M, Pruitt KD, Schuler GD, Bittner ML, Chen Y, Simon R, Meltzer P, Trent JM, Bogoski MS. 1998. Data management and analysis for gene expression arrays. Nat Genet 20:19–23.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiori MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer Class discovery and class prediction by gene expression monitoring. Science 286:531–537.

Li C, Wong WH. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc Natl Acad Sci USA 98:31–36.

Li C, Wong WH. 2001. Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. Genome Biology 2(8): 0032.1–0032.11.

Lipschutz RJ, Fodor S, Gingeras T, Lockhart D. 1999. High density synthetic oligonucleotide arrays. Nat Genet 21: 20–24.

Lockhart DJ, Dong H, Byrne M, Follettie M, Gallo M, Chee M, Mittmann M, Wang C, Kobayashi M, Horton H, Brown E. 1996. Expression monitoring by hyrbridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680.

Mahadevappa M, Wodicka L. 1999. A high-density probe array sample preparation method using 10- to 100-fold fewer cells. Nat Biotechnol 17:1134–1136.

Schadt EE, Li C, Su C, Wong WH. 2001. Analyzing high-density oligonucleotide gene expression array data. J Cell Biochem 80:192–202.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467–470.

Shalon D, Smith SJ, Brown PO. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6:639–645.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96:2907–2912.

Wahba G. 1990. Spline methods for observational data. CBMS-NSF regional conference series in applied mathematics. Philadelphia: SIAM.

Wodicka L, Dong H, Mittman H, Ho M, Lockhart D. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat Biotechnol 15: 1359–1366.

Zhu H, Cong JP, Mamtora G, Gingeras T, Shenk T. 1998. Cellular gene expression altered by human cytomegalovirus: Global monitoring with oligonucleotide arrays. Proc Natl Acad Sci USA 95:14470–14475.