

Statistical Analysis of Array Expression Data as Applied to the Problem of Tamoxifen Resistance

Susan G. Hilsenbeck, William E. Friedrichs, Rachel Schiff, Peter O'Connell, Rhonda K. Hansen, C. Kent Osborne, Suzanne A. W. Fuqua

Background: Although the emerging complementary DNA (cDNA) array technology holds great promise to discern complex patterns of gene expression, its novelty means that there are no well-established standards to guide analysis and interpretation of the data that it produces. We have used preliminary data generated with the CLONTECH Atlas™ human cDNA array to develop a practical approach to the statistical analysis of these data by studying changes in gene expression during the development of acquired tamoxifen resistance in breast cancer. **Methods:** For hybridization to the array, we prepared RNA from MCF-7 human breast cell tumors, isolated from our athymic nude mouse xenograft model of acquired tamoxifen resistance during estrogen-stimulated, tamoxifen-sensitive, and tamoxifen-resistant growth. Principal components analysis was used to identify genes with altered expression. **Results and Conclusions:** Principal components analysis yielded three principal components that are interpreted as 1) the average level of gene expression, 2) the difference between estrogen-stimulated gene expression and the average of tamoxifen-sensitive and tamoxifen-resistant gene expression, and 3) the difference between tamoxifen-sensitive and tamoxifen-resistant gene expression. A bivariate (second and third principal components) 99% prediction region was used to identify outlier genes that exhibit altered expression. Two representative outlier genes, *erk-2* and *HSF-1* (heat shock transcription factor-1), were chosen for confirmatory study, and their predicted relative expression levels were confirmed in western blot analysis, suggesting that semiquantitative

estimates are possible with array technology. **Implications:** Principal components analysis provides a useful and practical method to analyze gene expression data from a cDNA array. The method can identify broad patterns of expression alteration and, based on a small simulation study, will likely provide reasonable power to detect moderate-sized alterations in clinically relevant genes. [J Natl Cancer Inst 1999; 91:453-9]

Tremendous effort in cancer research has been devoted to identifying biologically relevant, differentially expressed genes by comparing, for example, tumor cells with normal cells or primary cells with metastatic cells. Until recently, most studies have been limited to quantitation of expression of at most a few genes at a time. Complementary DNA (cDNA) arrays offer the potential to simultaneously quantify expression of many genes. Advances in cDNA array technology to address issues, such as array size, probe density, probe content, and readout, now make this technology sufficiently flexible, accessible, and practical for application in the laboratory (1). The novelty of this technology means that there are no well-established and widely accepted standards to guide analysis and interpretation of the data that it produces. Thus far, cDNA arrays of one type or another have been most often used in paired comparisons (e.g., control versus cancer) to identify differentially expressed genes in only a few types of cancer, such as melanoma (2), Ewing's sarcoma (3), oral cancer (4), glioblastoma multiforme tumors (5), and gastrointestinal tumors (6). After standardization, rules for gene selection were typically based on ratios of expression [for example, greater than twofold difference (7), greater than three standard deviations of control genes ratio (2), or an arbitrary percent]. Application of the technology to more complex experimental designs involving simultaneous analysis of multiple experimental conditions or sampling over several time points will require a more general approach.

Tamoxifen is the most frequently prescribed drug for the treatment of breast cancer. Its use in breast cancer treatment has expanded from first-line treatment for

advanced metastatic disease (8), to adjuvant therapy after surgery for primary disease (9), and possibly to prevent breast cancer (10). Acquired tamoxifen resistance is a clinically important problem because a majority of patients with breast cancer will be offered tamoxifen at some time during their treatment, and although tamoxifen is initially effective in many patients, resistance eventually develops. Clinical resistance is almost certainly heterogeneous and multifactorial. Changes may be at the level of the target estrogen receptor (11-14), at a postreceptor point in the estrogen-receptor-response pathway (15-18), and/or downstream of the response pathway (19-21). With cDNA array technology (6,22,23), we may be able to discern the potentially complex patterns of gene expression that are involved in the acquisition of resistance.

In this study, we have used principal components analysis as a practical, but statistically valid, approach to simultaneously examine array data from several time points in an *in vivo* model of acquired resistance. The model simulates the clinical tamoxifen-resistant phenotype by using estrogen receptor-positive MCF-7 breast cancer tumors growing in athymic nude mice (24). We demonstrate that principal components analysis can reliably detect moderately sized alterations in gene expression that we have confirmed by western blot analysis.

MATERIALS AND METHODS

Tumors and Microarray Hybridization

MCF-7 breast cancer cells were injected into the mammary fat pads of athymic nude mice supplemented with an estrogen pellet as described previously (24) until tumors grew. The estrogen pellets were removed and the animals were treated with

Affiliations of authors: S. G. Hilsenbeck, W. E. Friedrichs, R. Schiff, R. K. Hansen, C. K. Osborne, S. A. W. Fuqua (Departments of Medicine/Oncology), P. O'Connell (Department of Pathology), The University of Texas Health Science Center, San Antonio.

Correspondence to: Suzanne A. W. Fuqua, Ph.D., The University of Texas Health Science Center, Departments of Medicine/Oncology, 7703 Floyd Curl Dr., San Antonio, TX 78248-7884 (e-mail: suzanne_fuqua@oncology.uthscsa.edu).

See "Notes" following "References."

© Oxford University Press

tamoxifen. Tumor volumes then declined and remained stable for several months. Invariably, however, after initial growth suppression, the tumors became resistant and growth resumed. Animals were killed at various times to obtain estrogen-stimulated tumors before tamoxifen treatment, tamoxifen-sensitive tumors during tamoxifen treatment but before acquired resistance, and tamoxifen-resistant tumors after tumor growth had resumed. We collected five tumors from each group. We then prepared total RNA with RNeasy kits (Qiagen Inc., Valencia, CA), and isolated messenger RNA on Dynabeads (Dyna, Oslo, Norway) according to manufacturer's instructions. For each group, the RNAs were pooled and used to synthesize ³²P-radiolabeled cDNAs for hybridization to the Atlas™ human cDNA expression array-1, according to the manufacturer's instructions (25) with SuperScriptII reverse transcriptase (Life Technologies, Inc. [Gibco BRL], Gaithersburg, MD). The CLONTECH Atlas™ human cDNA expression array is a positively charged nylon membrane (8 × 12 cm) that is spotted in duplicate with 200- to 600-base-pair cDNA fragments representing 588 genes and 21 housekeeping genes or control sequences (25). Genes are arrayed in six quadrants with genes of like function (i.e., oncogenes, assorted receptors, etc.) grouped together geographically. The hybridization data were collected with a Molecular Dynamics PhosphoImager™ (Molecular Dynamics, Sunnyvale, CA). This array was essentially the only one available when these experiments were done. Although the array does not include the estrogen receptor, it does include many other genes of potential interest in breast cancer, including two that we have studied previously, hsp27 and heregulin- α . We collected data from three arrays, one array for each tumor type.

Western Blot Analysis

Pulverized frozen tumors were manually homogenized in 5% sodium dodecyl sulfate. After boiling and microcentrifugation (10 minutes at 10 000 rpm, room temperature), clear supernatants were collected, and the protein concentration was determined by the bicinchoninic acid method (Pierce Chemical Co., Rockford, IL) as previously described (26). Twenty-five micrograms of protein was separated on a denaturing polyacrylamide gel and transferred by electroblotting to nitrocellulose membranes (Schleicher and Schuell, Inc., Keene, NH). The blots were first stained with StainAll dye (Alpha Diagnostic Intl., Inc., San Antonio, TX), to confirm uniform transfer of all samples, and then incubated in blocking solution (5% nonfat dry milk in Tris-HCl buffered saline-Tween [TBST = 50 mM Tris-HCl at pH 7.5, 150 mM NaCl, and 0.05% Tween 20]). After brief washes with TBST, the filters then were reacted with primary antibodies to erk-2 (UBI, Lake Placid, NY) or heat shock transcription factor-1 (HSF-1) (Stressgen, Victoria, Canada) for 1 hour at room temperature followed by extensive washes with TBST. Blots were then incubated with horseradish peroxidase-conjugated secondary antibody (Amersham Life Science Inc., Arlington Heights, IL) for 1 hour, washed with TBST, and developed by the ECL procedure (Amersham Life Science Inc.). The autoradiograms from the western blots were scanned with a densitometer, and the data are presented as the area determined for each individual tumor sample.

Statistical Considerations

In this pilot study, each hybridization ($m =$ three arrays) resulted in expression values for 588 genes and 21 control genes (putative housekeeping genes and negative control genes). The control genes, which were arrayed in a separate row at the bottom of the array and were more difficult to quantitate reliably in replicated experiments using the same RNA (data not shown), were not included in the statistical analyses. Expression of the highest and lowest expressed genes on the array varied by two to three orders of magnitude. Logarithmic transformation of the raw data reduced this range and helped equalize variability. This also means that additive effects on the log scale can be interpreted as fold changes in actual expression.

Because of the expense, limited amounts of RNA, and other considerations, array experiments usually have few replications and invariably have orders of magnitude more variables (genes and expressed sequence tags) than observations (hybridizations). In this study, we switch the roles of variables and observations, treating each tumor type as a variable ($m =$ three arrays) and each expressed gene sequence as an observation ($n =$ 588 genes).

Principal components analysis of mean-centered log-transformed data, based on the variance-covariance matrix (27), was then used to standardize across the three hybridizations and to extract three new axes (components P1, P2, and P3), expressed as linear combinations of the original axes (variables ES [estrogen-stimulated], TS [tamoxifen-sensitive], and TR [tamoxifen-resistant]).

$$P1 = A_1 * ES + B_1 * TS + C_1 * TR$$

$$P2 = A_2 * ES + B_2 * TS + C_2 * TR$$

$$P3 = A_3 * ES + B_3 * TS + C_3 * TR$$

In principal components analysis, the coefficients (As, Bs, and Cs) are chosen so that the first component (P1) explains the maximal amount of variance in the data. The second component (P2) is perpendicular to the first and explains the maximal residual squared variation, and the third component (P3) is perpendicular to the first two. Meaning was ascribed to the new axes by visual examination of the coefficients. In these array experiments, P1 represents the average level of expression across the tumor types and P2 and P3 represent differences between tumor types. A bivariate analysis that results in two new axes (P1 and P2) was also performed to compare tamoxifen-sensitive gene expression with tamoxifen-resistant gene expression. The coefficients do not always have a nice biologically sensible interpretation, although the higher-order components can still be used to identify outlier genes, regardless of interpretation (see below).

We used P2 (and P3 in the higher-order analysis) to identify outlier genes that might represent true alterations in gene expression. In the bivariate principal components analysis of tamoxifen-sensitive gene expression versus tamoxifen-resistant gene expression, we used a normal approximation to construct a 99% prediction region for component P2 (i.e., $0 \pm 2.57 * SD_r$, where $SD_r =$ interquartile range/1.35). A robust estimate of the standard deviation (SD_r) was used to reduce the variance-inflating effects of outliers (28). Genes outside the region were identified for further study. Analo-

gously, in a trivariate principal components analysis (estrogen-stimulated, tamoxifen-sensitive, and tamoxifen-resistant gene expression), we computed a 99% bivariate normal prediction ellipse (27,29) for components P2 versus P3, and genes outside the ellipse were selected for investigation.

This "robust prediction interval" approach seems justified on the following basis. Although the distribution of P1 is highly skewed, higher-order components are roughly symmetric. When there is no differential expression, as in a bivariate analysis of two array hybridizations using the same pool of RNA, the higher-order components are approximately normally distributed (data not shown). In experiments comparing different pools of RNA, where some genes may be differentially expressed, the observed distribution of each higher-order component (P2, P3, etc.) should be a mixture of central ($\mu = 0$) and noncentral ($\mu \neq 0$) distributions. By using a robust estimator that focuses on the middle of the observed distribution, which should represent primarily unaltered genes, we hope to increase sensitivity to identify truly altered genes. The prediction level (99%), which is analogous to the specificity of a diagnostic test, was chosen arbitrarily as representing a reasonable balance between identifying too many spuriously "significant" genes and missing true alterations. For display purposes, we have back-transformed the data by exponentiation of P2 and P3 so that the data are shown as approximate fold increases or decreases in expression.

The ability of this methodology to detect true alterations was examined in a small simulation study. Log-transformed values from a hypothetical bivariate array experiment with 588 genes were generated to have a common log-normally distributed component for level of expression [i.e., $\exp(X) + 8$, where $X \sim N(\mu = 0, \sigma = 0.6)$], and independent normally distributed errors [i.e., $\log_e(\text{Control}) = \exp(X) + 8 + Y$ and $\log_e(\text{Experimental}) = \exp(X) + 8 + Z$, where $Y, Z \sim N(\mu = 0, \sigma = 0.17)$].

The distributional parameters were chosen to mimic data seen in our real experiments. A small percentage of truly altered genes (2% or 4%) were created by shifting the error distribution for the experimental member of the pair up or down (with 50% probability) to represent an average 2- or 2.5-fold change from baseline [i.e., $\log_e(\text{Experimental}) = \exp(X) + 8 + W$, where $W \sim N(\mu = \pm 0.7, \sigma = 0.17)$]. The generated data were then analyzed as described above, and the numbers of truly altered and spuriously altered genes falling outside the 99% prediction region were tabulated. Each scenario was replicated 100 times, and the results were summarized over all replications. All analyses were performed with the SAS program package (Version 6.11, SAS Institute, Cary, NC).

RESULTS

Bivariate Analysis

Fig. 1 shows the three bivariate log-log scatter plots that arise from pairwise comparisons of the data from the three cDNA array hybridizations (one for estrogen-stimulated tumors, one for tamoxifen-sensitive tumors, and one for tamoxifen-resistant tumors). Each gene of the 588 genes on the array (excluding housekeep-

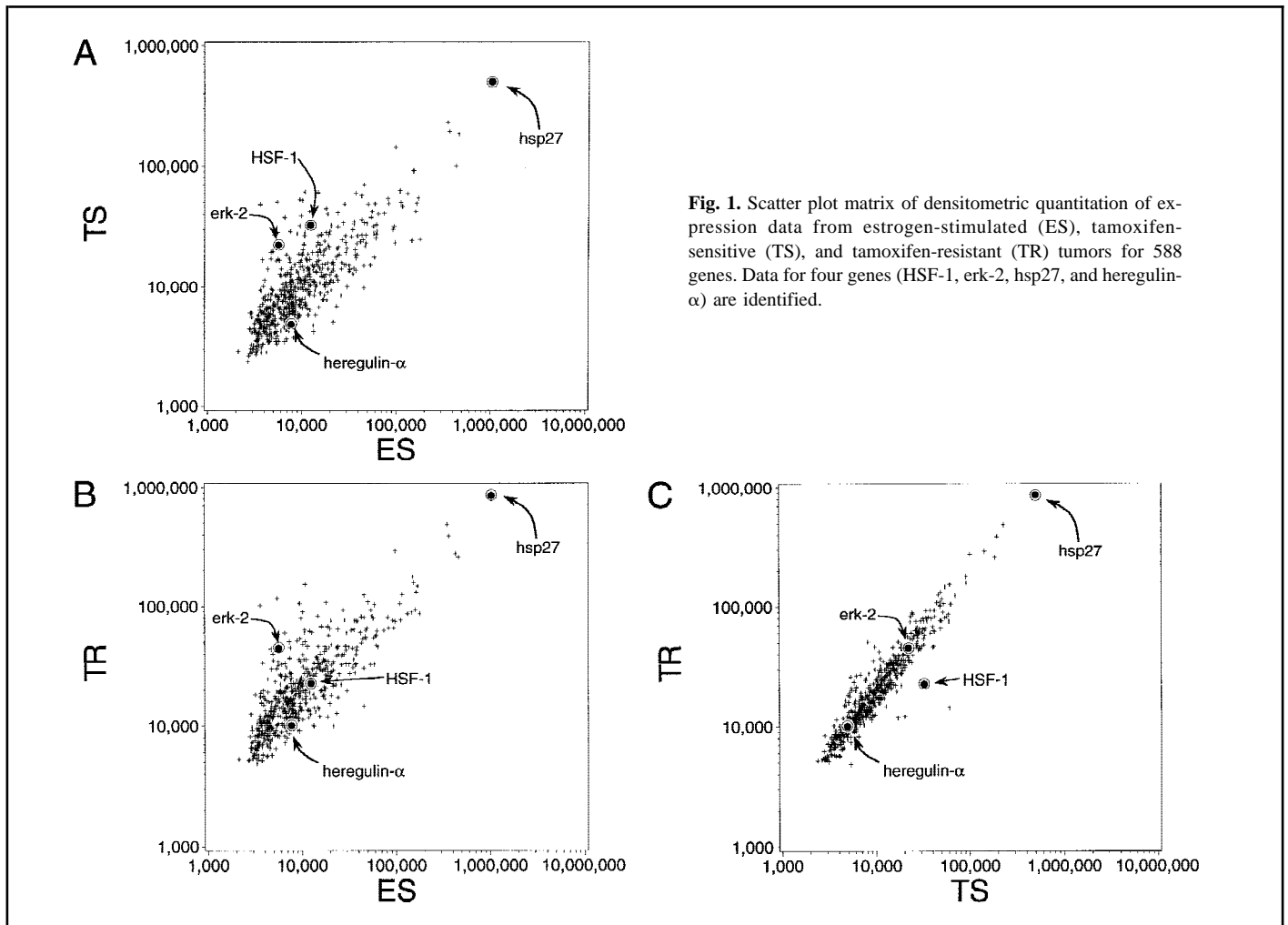


Fig. 1. Scatter plot matrix of densitometric quantitation of expression data from estrogen-stimulated (ES), tamoxifen-sensitive (TS), and tamoxifen-resistant (TR) tumors for 588 genes. Data for four genes (HSF-1, erk-2, hsp27, and heregulin- α) are identified.

ing and control genes) is represented by a point on the scatter plots. The individual values ranged over two to three orders of magnitude, indicating that the most highly expressed genes were expressed at 100- or 1000-fold higher levels than the lowest expressed genes. For example, the 27-kd heat shock protein (hsp27) was the most highly expressed gene on the array in all three tumor types. This finding is consistent with our previously published result that hsp27 is amplified and overexpressed in the late-passage MCF-7 cells used in this model (30). Similarly, the array results are consistent with previous findings (31) that heregulin- α is expressed at relatively low levels in all three types of tumor cells.

In each scatter plot, most genes lie fairly close to a diagonal line of "identity." This line may not be centered on the graph if there are differences in the average level of radioactivity of probes used in each hybridization. The distance along this line denotes differences in the level of expression between genes, such as we see

between hsp27 and heregulin- α . The perpendicular distance away from the line denotes differences in expression within the same gene between tumor types.

Principal components analysis of the log-transformed expression data was used to produce a new set of axes (Fig. 2). For tamoxifen-sensitive tumors versus tamoxifen-resistant tumors (Fig. 2, A), the new x axis or first principal component (P1) roughly corresponds to the line of "identity" and represents level of expression. The second principal component (P2) is perpendicular to the first and represents difference in expression between tumor types. In the bivariate analysis, more than 97% of the total variation in the log-transformed data was associated with P1, leaving about 3% for P2. The two components are, by definition, not correlated ($\rho = 0$). The distribution of P1 is skewed, because many genes on the array are expressed at low to moderate levels, but only a few are expressed at extremely high levels. The distribution of P2 is roughly symmetric, and a 99% robust pre-

diction interval identified 35 outlier genes that may be over- or under-expressed in tamoxifen-resistant tumors relative to tamoxifen-sensitive tumors (Fig. 2, B).

Trivariate Analysis

Bivariate principal components analysis could be performed for each pair of tumor types; however, a more comprehensive three-way analysis is preferred and is more biologically relevant. Principal components analysis of the mean-centered log-transformed data (for estrogen-stimulated tumors, tamoxifen-sensitive tumors, and tamoxifen-resistant tumors) yields three new axes (P1, P2, and P3) that account for 90.5%, 8%, and 1.5% of the variation in the data, respectively. By inspection of the coefficients, the first principal component (P1) is again interpreted as the "average level of expression" because the coefficients were all positive and similar in value (0.63, 0.55, and 0.55, respectively). The second principal component (P2) clearly contrasts data from estrogen-stimulated tu-

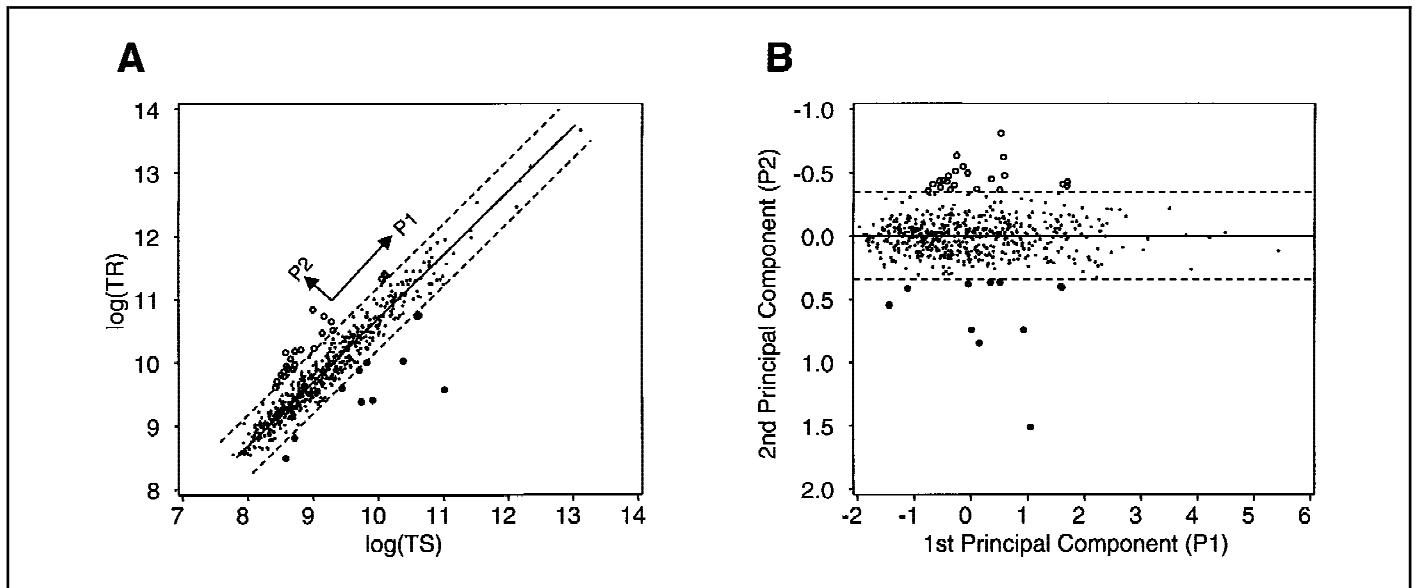


Fig. 2. A) Scatter plot of log-transformed expression data for tamoxifen-sensitive (TS) and tamoxifen-resistant (TR) tumors showing the line of identity (solid line) and 99% prediction region (dashed line). The open and solid circles indicate genes that are overexpressed or underexpressed, respectively, in tamoxifen-resistant tumors relative to tamoxifen-sensitive tumors. Dots indicate genes that are not affected by tamoxifen. **B)** Scatter plot of first and second principal components from the same data shown in A.

mors to the average of tamoxifen-sensitive and tamoxifen-resistant tumors because the P2 coefficient for the estrogen-stimulated data is negative (-0.78) and roughly equal to the sum of the tamoxifen-sensitive and tamoxifen-resistant coefficients (0.46 and 0.43 , respectively). The third principal component (P3) primarily represents differences between the tamoxifen-sensitive and the tamoxifen-resistant tumors, because the P3 coefficient for the estrogen-stimulated tumors is small (0.02) and the tamoxifen-sensitive and tamoxifen-resistant coefficients are nearly equal but opposite in sign (0.69 and -0.72 , respectively). Fig. 3 shows a scatter plot of P2 versus P3. Points near the center represent genes that were similarly expressed in all three tumor types, whereas points on the periphery exhibit alterations in expression. Data have been back-transformed to show the approximate fold changes in expression. We used a bivariate normal approximation with robust estimates of standard deviations to compute a 99% prediction ellipse. Genes lying outside the region may exhibit real alterations in the level of expression that are associated with the biologic effects during the transition from estrogen-stimulated to tamoxifen-sensitive status and tamoxifen-sensitive to tamoxifen-resistant status.

In addition, different regions of the P2 \times P3 plane correspond to different temporal patterns of expression alteration. For

example, expression of genes to the far right in Fig. 3 (i.e., near *erk-2*) is increased by tamoxifen relative to the expression of genes in estrogen-stimulated tumors but expression of genes in this area is unchanged in tamoxifen-resistant tumors relative to tamoxifen-sensitive tumors. In contrast, expression of genes to the lower right in Fig. 3 (i.e., near *HSF-1*) is increased in tamoxifen-sensitive tumors relative to estrogen-stimulated tumors but is decreased in tamoxifen-resistant tumors.

Confirmation of Gene Expression by Western Blot Analysis

We selected two genes just outside of the 99% prediction ellipse (*erk-2* and *HSF-1*) for quantitation by western blot analysis. These two genes were chosen because of their relatively low expression (Fig. 1) and modest alteration, so that we could address sensitivity questions and the ready availability of specific antibodies. The *erk-2* kinase is a known mediator of the growth factor signaling pathway, and it has been shown that the estrogen receptor can activate its activity in MCF-7 cells (32). *HSF-1* is involved in cellular stress responses (33) and is thus a potential marker of tamoxifen-induced stress. We found that the relative levels of *erk-2* and *HSF-1* predicted in the array experiment were indeed confirmed in an independent set of individual tumors (Fig. 3,

B, lanes 1–15) from the athymic nude mouse model. As predicted by Figs. 1, A, and 3, A, western blot results for *HSF-1* indicate a substantial increase in expression in tamoxifen-sensitive tumors relative to estrogen-stimulated tumors, which is followed by a decrease in tamoxifen-resistant tumors to approximately the levels in estrogen-stimulated tumors (Fig. 1, B). Similarly for *erk-2*, there is an increase in expression in tamoxifen-sensitive tumors relative to estrogen-stimulated tumors (Fig. 1, A), but there is relatively less change between tamoxifen-sensitive and tamoxifen-resistant tumors.

Power Considerations

Using distributional parameters from some of our pilot studies, we ran a series of simulations to investigate the likely sensitivity of these methods to detect real differences of moderate size (Table 1). With modest changes (twofold) in 2%–4% of genes, 99% of the unchanged genes were correctly classified as unchanged by the 99% prediction interval, and 59% of the altered genes were correctly identified as outliers. With larger differences (e.g., 2.5-fold), the proportion of correctly identified outliers goes up (85%). Although the outliers will always be contaminated by a few spuriously identified genes, these results suggest that the method has reasonable power to detect real differences.

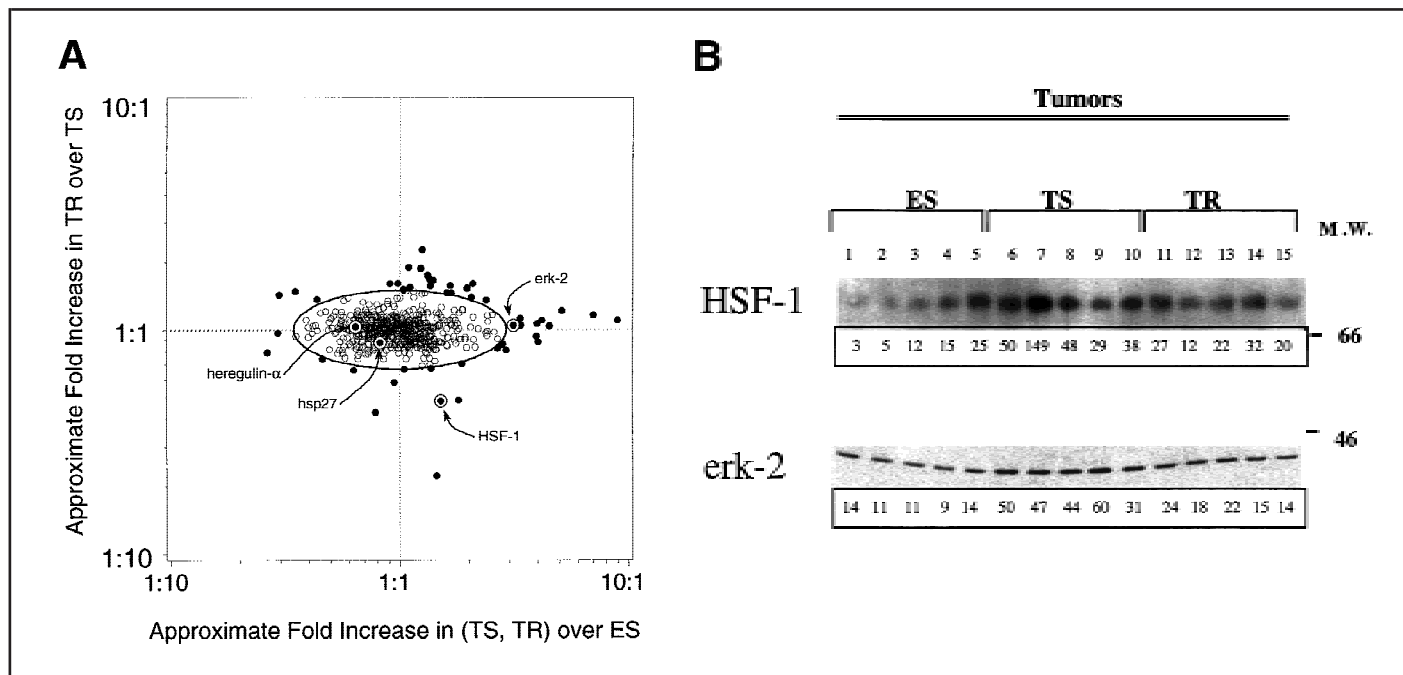


Fig. 3. A) Scatter plot of second and third principal components from principal components analysis of log-transformed gene expression data from estrogen-stimulated (ES), tamoxifen-sensitive (TS), and tamoxifen-resistant (TR) tumors, back-transformed to show approximate fold alterations. Axis labels describe the qualitative interpretation of principal components analysis coefficients. Genes inside or outside the 99% prediction ellipse (solid line) are shown as open or solid circles, respectively. Data for four genes (HSF-1, erk-2, hsp27, and heregulin- α) are identified. **B)** Western blot analysis with erk-2 and HSF-1 antibodies in estrogen-stimulated (ES, lanes 1–5), tamoxifen-sensitive (TS, lanes 6–10), and tamoxifen-resistant (TR, lanes 11–15) tumors (five tumors are in each

group). The positions of molecular weight (M.W.) markers (in 10^{-3} kd) are shown to the right. Densitometric scan values (in relative units) for each lane are shown in the boxed area below each western blot lane. For HSF-1 protein expression, there was a fivefold increase in the tamoxifen-sensitive tumors and a 1.8-fold increase in the tamoxifen-resistant tumors relative to the estrogen-stimulated tumor group. For erk-2 expression, there was a fourfold increase in the tamoxifen-sensitive tumors and a 1.6-fold increase in the tamoxifen-resistant tumors relative to the estrogen-stimulated tumor group; the slight difference in protein levels compared with that predicted by the RNA array analysis may reflect posttranscriptional and/or translational control of erk-2 protein.

Table 1. Results of simulation study involving 588 genes in two tumor types and using a 99% prediction interval

Average fold change in altered gene expression	% of genes with altered expression	% of genes with unaltered expression inside interval*	% of genes with altered expression outside interval†
2.0	2	99	59
	4	99	60
2.5	2	99	86
	4	99	85

*This is the observed specificity and is analogous to prediction level $(1 - \alpha)$.

†This is the observed sensitivity and is analogous to power.

DISCUSSION

cDNA microarray expression profiling offers tremendous potential to simultaneously characterize the expression of large numbers of gene sequences. In theory, comparisons of hybridization data from pairs or a series of RNA pools, representing cells from various tumors or experimental conditions, should allow us to identify differentially expressed genes or sequences that may be involved in the biologic process under investigation. In practice, it is not so easy to distinguish true differences in expression from differences in expression due to experimental

variability only. In a traditional study of one or a few genes, statistical analysis of experimental replicates would be used to estimate variability in expression for each gene to determine whether expression is altered. Variability between replicates is often large, and moderate-sized differences (two- to 10-fold) can require many experimental replications. Due to expense, limited amounts of RNA, and other considerations, array experiments usually have few replications and invariably have orders of magnitude more variables (genes and expressed sequence tags) than observations. In our study of acquired tamoxifen resistance, we have switched

the roles of variables and observations and used principal components analysis, coupled with robust estimates of 99% prediction regions on higher-order components, as a practical approach to screening array data for likely candidates for further study. The method presumes that the vast majority of genes will be altered very little and uses information from all genes to obtain more stable estimates of variability. The method is not limited to pairwise comparisons but can be used to study several tumor types or experimental conditions simultaneously. In a small simulation study, we have shown that this approach is capable of reliably identifying 60%–85% of genes exhibiting moderate degrees of differential expression (2- to 2.5-fold), without increasing the number of spuriously identified outliers.

In this study, we used an *in vivo* athymic mouse model of acquired tamoxifen resistance (24) to explore the power of microarray expression profiling. In this tamoxifen-resistance model, we have previously shown that one potential resistance mechanism is stimulation of the tumor by tamoxifen, which acts as a partial agonist. As our first analysis, we used the

array technology to identify those genes that might be associated with this growth stimulation. We hypothesized that the tamoxifen-stimulated phenotype could result from the deregulated expression of downstream growth-regulatory pathways that liberate the cell cycle from normal steroid control. Indeed, it has been reported that overexpression of single growth regulatory genes such as cyclin D1 (34), protein kinase A (35), and transforming growth factor β (21) can influence a cell's response to tamoxifen treatment. However, there are probably multiple mechanisms that coexist in tumors and in conjunction contribute to the clinical tamoxifen-resistant phenotype. The microarray expression profiling technology is well-suited for this clinical problem. Principal components analysis of our preliminary data suggests that distinct patterns of temporal alteration in gene expression can be distinguished. Our future studies will be aimed at identifying which of the outlier genes are most contributory to the tamoxifen-stimulated phenotype and testing these genes in clinical samples on custom microarrays. From these studies, we expect to identify the gene expression patterns predictive of tamoxifen-resistant growth.

In summary, principal components analysis of log-transformed array data provides a practical approach to data reduction, visualization, and identification of "significant" outlier genes. As a result, analysis of cDNA expression arrays can identify genes and pathways that are altered during the process of resistance. We predict that principal components analysis or related methods of analysis of microarray expression data will lead to the identification of novel growth pathways that are important for the generation of tamoxifen resistance and thus will generate new predictive clinical paradigms.

REFERENCES

- (1) Marshall A, Hodgson J. DNA chips: an array of possibilities. *Nat Biotechnol* 1998;16:27-31.
- (2) DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyze gene expression patterns in human cancer. *Nat Genet* 1996;14:457-60.
- (3) Welford SM, Gregg J, Chen E, Garrison D, Sorensen PH, Denny CT, et al. Detection of differentially expressed genes in primary tumor tissues using representational differences analysis coupled to microarray hybridization. *Nucleic Acids Res* 1998;26:3059-65.
- (4) Chang DD, Park NH, Denny CT, Nelson SF, Pe M. Characterization of transformation related genes in oral cancer cells. *Oncogene* 1998;16:1921-30.
- (5) Sehgal A, Boynton AL, Young RF, Vermeulen SS, Yonemura KS, Kohler EP, et al. Application of the differential hybridization of Atlas Human expression arrays technique in the identification of differentially expressed genes in human glioblastoma multiforme tumor tissue. *J Surg Oncol* 1998;67:234-41.
- (6) Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, et al. Gene expression profiles in normal and cancer cells. *Science* 1997;276:1268-72.
- (7) Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A* 1996;93:10614-9.
- (8) Bezwoda WR, Esser JD, Dansey R, Kessel I, Lange M. The value of estrogen and progesterone receptor determinations in advanced breast cancer. Estrogen receptor level but not progesterone receptor level correlates with response to tamoxifen. *Cancer* 1991;68:867-72.
- (9) Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. Early Breast Cancer Trialists' Collaborative Group. *Lancet* 1992;339:71-85.
- (10) Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel P-1 Study. *J Natl Cancer Inst* 1998;90:1371-88.
- (11) Fuqua SA, Fitzgerald SD, Chamness GC, Tandon AK, McDonnell DP, Nawaz Z, et al. Variant human breast tumor estrogen receptor with constitutive transcriptional activity. *Cancer Res* 1991;51:105-9.
- (12) Fuqua SA, Wiltshcke C, Castles C, Wolf D, Allred DC. A role for estrogen receptor variants in endocrine resistance. *Endocrine-related cancer* 1995;2:19-25.
- (13) Daffada AA, Johnston SR, Smith IE, Detre S, King N, Dowsett M. Exon 5 deletion variant estrogen receptor messenger RNA expression in relation to tamoxifen resistance and progesterone receptor/pS2 status in human breast cancer. *Cancer Res* 1995;55:288-93.
- (14) Gallacchi P, Schoumacher F, Eppenberger-Castori S, Von Landenberg EM, Kueng W, Eppenberger U, et al. Increased expression of estrogen-receptor exon-5-deletion variant in relapse tissues of human breast cancer. *Int J Cancer* 1998;79:44-8.
- (15) Smith CL, Nawaz Z, O'Malley BW. Coactivator and corepressor regulation of the agonist/antagonist activity of the mixed antiestrogen, 4-hydroxytamoxifen. *Mol Endocrinol* 1997;11:657-66.
- (16) Jackson TA, Richer JK, Bain DL, Takimoto GS, Tung L, Horwitz KB. The partial agonist activity of antagonist-occupied steroid receptors is controlled by a novel hinge domain-binding coactivator L7/SPA and the corepressors N/CoR or SMRT. *Mol Endocrinol* 1997;11:693-705.
- (17) Lavinsky RM, Jepsen K, Heinzel T, Torchia J, Mullen TM, Schiff R, et al. Diverse signaling pathways modulate nuclear receptor recruitment of N-CoR and SMRT complexes. *Proc Natl Acad Sci U S A* 1998;95:2920-5.
- (18) Berns EM, van Staveren IL, Klijn JG, Foekens JA. Predictive value of SRC-1 for tamoxifen response of recurrent breast cancer. *Breast Cancer Res Treat* 1998;48:87-92.
- (19) Zwijsen RM, Wientjens E, Klompmaaker R, van der Sman J, Bernards R, Michalides RJ. CDK-independent activation of estrogen receptor by cyclin D1. *Cell* 1997;88:405-15.
- (20) Lonning E, Lien EA. Mechanisms of action of endocrine treatment in breast cancer. *Crit Rev Oncol Hematol* 1995;21:158-93.
- (21) Thompson AM, Kerr DJ, Steel CM. Transforming growth factor beta 1 is implicated in the failure of tamoxifen therapy in human breast cancer. *Br J Cancer* 1991;63:609-14.
- (22) Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;270:467-70.
- (23) Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, et al. Accessing genetic information with high-density DNA arrays. *Science* 1996;274:610-4.
- (24) Osborne CK, Hobbs K, Clark GM. Effect of estrogens and antiestrogens on growth of human breast cancer cells in athymic nude mice. *Cancer Res* 1985;45:584-90.
- (25) CLONTECH Laboratories Inc. Atlas™ Human cDNA expression arrays user manual. Palo Alto (CA): CLONTECH Laboratories, Inc; 1997.
- (26) Tandon AK, Clark GM, Chamness GC, Ullrich A, McGuire WL. HER-2/neu oncogene protein and prognosis in breast cancer. *J Clin Oncol* 1989;7:1120-8.
- (27) Tatsuoaka MM, editor. *Multivariate analysis: techniques for educational and psychological research*. New York (NY): John Wiley & Sons, Inc; 1971. p. 94-149.
- (28) Venables WN, Ripley BD. *Modern applied statistics with S-plus*. New York (NY): Springer-Verlag; 1994. p. 203-8.
- (29) Anderson TW. *An introduction to multivariate statistical analysis*. New York (NY): John Wiley & Sons, Inc.; 1958. p. 123.
- (30) Fuqua SA, Benedix MG, Krieg S, Weng CN, Chamness GC, Oesterreich S. Constitutive overexpression of the 27,000 dalton heat shock protein in late passage human breast cancer cells. *Breast Cancer Res Treat* 1994;32:177-86.
- (31) Tang CK, Perez C, Grunt T, Waibel C, Cho C, Lupu R. Involvement of heregulin-beta2 in the acquisition of the hormone-independent phenotype of breast cancer cells. *Cancer Res* 1996;56:3350-8.
- (32) Migliaccio A, Di Domenico M, Castoria G, de Falco A, Bontempo P, Nola E, et al. Tyrosine

- kinase/p21ras/MAP-kinase pathway activation by estradiol-receptor complex in MCF-7 cells. *EMBO J* 1996;15:1292-300.
- (33) Rabindran SK, Giorgi G, Clos J, Wu C. Molecular cloning and expression of a human heat shock factor, HSF1. *Proc Natl Acad Sci U S A* 1991;88:6906-10.
- (34) Neuman E, Ladha MH, Lin N, Upton TM, Miller SJ, DiRenzo J, et al. Cyclin D1 stimulation of estrogen receptor transcriptional activity independent of cdk4. *Mol Cell Biol* 1997;17:5338-47.
- (35) Fujimoto N, Katzenellenbogen BS. Alteration in the agonist/antagonist balance of antiestrogens by activation of protein kinase A signaling pathways in breast cancer cells: antiestrogen selectivity and promoter dependence. *Mol Endocrinol* 1994;8:296-304.

NOTES

Supported by USAMRDC DAMD17-94-J-4112 and Public Health Service grants CA58183, CA30195, and CA54174, National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

We thank Julia Perkins for preparation of the manuscript.

Manuscript received August 3, 1998; revised December 18, 1998; accepted December 30, 1998.