Estrogen receptor alpha positive breast tumors and breast cancer cell lines share similarities in their transcriptome data structures

YUELIN ZHU[1], ANTAI WANG[1,2], MINETTA C. LIU[1], ALAN ZWART[1], RICHARD Y. LEE[1], ANN GALLAGHER[1], YUE WANG[4], WILLIAM R. MILLER[5], J. MICHAEL DIXON[5] and ROBERT CLARKE[1,3]

[1]Lombardi Comprehensive Cancer Center and Departments of Oncology, [2]Biostatistics, Bioinformatics and Biomathematics, and [3]Physiology and Biophysics, Georgetown University School of Medicine, Washington, DC; [4]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA; [5]Department of Oncology, University of Edinburgh, Western General Hospital, Crewe Rd South, Edinburgh, Scotland, UK

Received March 7, 2006; Accepted May 5, 2006

Abstract. Established human breast cancer cell lines are widely used as experimental models in breast cancer research. While these cell lines and their variants share many phenotypic characteristics with human breast tumors, the extent to which they reflect the underlying molecular biology of breast cancer remains controversial. We explored this issue using a probabilistic rather than heuristic approach. Data from gene expression microarrays were used to compare the global structures of the transcriptomes of three estrogen receptor alpha positive (ER^+) human breast cancer cell lines (MCF-7, T47D, ZR-75-1) and 13 human breast tumors (11 ER^+; 2 ER^-). Linear representations of the respective data structures were obtained by deriving those top principal components (PCs) required to capture ≥80% of the cumulative variance for each data set (M PCs). We then identified those genes most highly correlated with the M PCs (Pearson's correlation coefficient $r \geq 0.800$) and identified a group of 36 genes commonly correlated with both the cell line (M = 5 PCs) and tumor (M = 6 PCs) data structures. All 36 common genes were correlated with PC1 from the breast tumor data: 21/36 genes were correlated with PC1, 14/36 genes correlated with PC2, and 1/36 genes correlated with PC3 from the cell line data. Genes important in defining the data structures include NFκB p65, IGFBP-6, ornithine decarboxylase-1, and paxillin. When data from MDA-MB-435 xenografts (ER^-) were included in the analysis, we were unable to find any common genes between these xenografts and the breast tumors. These data clearly imply that MCF-7, T47D, and ZR-75-1 cells and ER^+ breast tumors share substantial global similarities in the structures of their respective transcriptomes, and that these cell lines are good models in which to identify molecular events that are likely to be important in some ER^+ human breast cancers.

Introduction

Human breast cancer cell lines, whether growing *in vitro* or *in vivo* as xenografts in immunedeficient rodents, are among the most widely used experimental models in breast cancer research (1-4). These cell lines and their variants have been particularly useful as experimental models and enable investigators to address hypotheses in ways that would be technically difficult or ethically inappropriate in humans. We and others have extensively reviewed the characteristics of selected human breast cancer cell lines, their phenotypes, and the extent to which these phenotypes reflect key components of the human disease (2-4).

Almost 100 breast cancer cell lines have been described but Lacroix and Leclercq estimated that over two-thirds of studies involved work with only one or more of three models (4): the estrogen receptor alpha positive (ER^+), estrogen-dependent and antiestrogen sensitive MCF-7 and T47D cell lines, and the ER^-, estrogen-independent and antiestrogen resistant MDA-MB-231 cell line (3). All three cell lines are tumorigenic and locally invasive in immunedeficient rodents (2) but only orthotopic xenografts of MDA-MB-231 cells produce spontaneous metastases (5). ZR-75-1 is another commonly used ER^+ breast cancer cell line and is phenotypically similar to MCF-7 and T47D cells (3). MDA-MB-435 cells are widely used as a

Correspondence to: Dr Robert Clarke, Department of Oncology, Room W405A Research Building, Georgetown University School of Medicine, 3970 Reservoir Rd, NW, Washington, DC 20057, USA
E-mail: clarker@georgetown.edu

Abbreviations: M, essential dimensionality; ER, estrogen receptor alpha; PC, principal component; PCA, principal component analysis; tgDNA, total genomic DNA

Key words: breast cancer, bioinformatics, cell lines, data structure, genomics, microarray analysis, principal component analysis

metastatic ER$^-$ model (5); while these cells are phenotypically similar to MDA-MB-231 cells, the breast origin of the MDA-MB-435 cell line has been questioned (6,7).

It is evident that human breast cancer cell lines reflect important phenotypic characteristics present in the human disease, and they have been central to discovering and extending new knowledge in many areas of breast cancer research (4). However, the extent to which biological insights can be extrapolated from preclinical models to the human disease remains somewhat controversial. Established breast cancer cell lines exhibit substantial aneuploidy and genetic instability, and variants can arise spontaneously over time (8). While this is probably a reflection of the inherent molecular/genetic instability of breast tumors, it is unclear how well human breast cancer cell lines growing *in vitro* reflect the underlying molecular biology of breast tumors. For example, a study comparing the molecular profiles of 60 human cell lines showed that, by unsupervised hierarchical clustering, breast cancer cell lines do not cluster together but are scattered across the entire dendrogram (6). These investigators also reported a hierarchical clustering analysis of data restricted to five breast cancer cell lines, four leukemia cell lines, two breast tumors, one breast tumor metastasis, and one specimen of normal breast tissue. The breast cancer cell lines clustered together but this cluster was more similar to the leukemia cell lines than to the breast tumors. Moreover, the normal breast specimen and the breast tumors formed a single cluster distinct from all the cell lines (6). A subsequent review of these and other molecular profiles concluded that breast cancer cell lines and tumors shared some gene expression patterns in common. However, the authors took a largely intuitive rather than probabilistic approach, looking for commonalities in gene expression patterns in cell lines with predetermined cellular phenotypes/functions. The authors acknowledged that alternative interpretations of the data were possible (9).

In this study, our primary goal was to obtain a relatively unbiased probabilistic assessment of the global similarities in the transcriptomes of human breast cancer cell lines and breast tumors. Rather than compare broadly defined phenotypic or genetic characteristics, we asked directly whether similarities exist within the structures of their respective high dimensional gene expression microarray data. To address this goal, we first developed an application of principal components analysis (PCA) (10) based on the general approach described by Jolliffe (11). PCA is a technique that finds linear transformations of data such that the first principal component (PC) is that linear projection that best captures the greatest variance in the data. The second PC is orthogonal to the first and captures the second greatest variance, and so on. In this manner, PCA can be used to find those projections that best capture the overall structure of the data. We show that three of the most widely used ER$^+$ human breast cancer cells lines (MCF-7, T47D, ZR-75-1) exhibit substantial similarities in their transcriptome data structures to a panel of mostly ER$^+$ breast cancer specimens from patients.

Materials and methods

Human breast cancer cell lines. MCF-7 cells were originally obtained from the Barbara A. Karmanos Cancer Institute (Detroit, MI), T47D and ZR-75-1 cells were obtained from the American Type Culture Collection (Manassas, VA), and MDA-MB-435 cells were from Dr Janet Price (M.D. Anderson Cancer Center, Houston, TX). All cell lines were maintained at 37˚C in cell culture medium (improved minimal essential medium with phenol red and supplemented with 5% (v/v) heat-inactivated fetal bovine serum; Biofluids, Rockville, MD) in a 95% air/5% CO_2 atmosphere. All cell lines were shown to be free of contamination with *Mycoplasma* spp.

Human breast tumor specimens. The 13 breast tumor specimens and the associated microarray data used in this study have been previously reported (12). Five of the 13 specimens were obtained from patients undergoing a diagnostic core needle or excisional biopsy at Georgetown University Hospital. All patients signed a written consent approved by the Georgetown University Medical Center Institutional Review Board. Core needle biopsies were either obtained under mammographic or ultrasound guidance during a routine diagnostic procedure, or obtained intraoperatively after surgical exposure of the tumor. The study pathologist performed a routine histopathologic analysis of frozen sections from all biopsies as previously described (12); biopsies were released for microarray analysis only if they did not contain any new clinical information important for patient care. The other eight breast tumor specimens were obtained at the Department of Oncology, University of Edinburgh (Scotland, UK); samples were collected with appropriate patient consent, and all procedures were performed using guidelines consistent with the relevant UK legislation. Once released for study, all patient identifiers were removed from each sample. Information not already published on these samples is included in Table I. The clinical material, mostly frozen in OCT, was directly provided to the research laboratory for storage and/or processing, whereupon tissue was either stored at -80ºC or processed immediately for RNA extraction.

MDA-MB-435 human breast cancer xenografts. Cells from subconfluent monolayers were removed by trypsinization. To establish xenografts, 1x10^6 viable cells, as estimated by trypan-blue dye exclusion, were subcutaneously inoculated into the region of the mammary fat pad as previously described (2,13). Mice were 4-6 week old female, NCr *nu/nu* athymic mice (~20 g body weight) and were housed 4 or 5 per cage and fed sterilized, pelleted food and sterilized water *ad libitum*. Nude mice (38) were used and tumors were observed at each of the inoculation sites. Tumors were measured twice weekly for 4 weeks *post inoculum*; consistently proliferating tumors were identified and removed immediately *post mortem* using sterile scissors and forceps. Studies were performed by the Lombardi Comprehensive Cancer Center Animal Research Shared Resource in a pathogen-free environment within a central facility approved by the American Association for Accreditation of Laboratory Animal Care. All work that required the use of vertebrate animals was performed in accordance with the current regulations and standards described by the United States Department of Agriculture and the United States Department of Health and Human Services, and with the approval of the Georgetown University Animal Care and Use Committee.

Table I. Characteristics of the breast tumor specimens.

Tumor	ER	Lymph nodes	% Cancer	Source
1	+	+	90	GU
2	+	-	80	GU
3	+	+	90	GU
4	+	+	90	GU
5	-	ND	80	GU
6	+	+	90	EU
7	+	+	90	EU
8	+	ND	99	EU
9	+	-	90	EU
10	+	+	90	EU
11	+	-	70	EU
12	+	+	90	EU
13	-	+	90	EU

ER, estrogen receptor alpha (positive, +; or negative, -); lymph nodes, presence (+), absence (-) of involved lymph nodes, or no data (ND); % cancer, proportion of each specimen that contains neoplastic breast epithelial cells; Source, center at which cases were accrued; GU, Georgetown University; EU, Edinburgh University. Additional information on selected cases has been previously published (12).

RNA preparation and gene expression microarray studies. Study materials were collected over a prolonged period and were processed slightly differently. These differences replicate some of the methodologic variability anticipated across laboratories but might be expected to introduce some noise into the data. For cell lines growing *in vitro*, each cell line sample represents data from an independent cell culture grown on a different day; no cultures were pooled, nor were RNAs extracted from cultures grown at the same time. Subconfluent monolayers were rapidly trypsinized, cells were centrifuged at 1,000 x g for 5 min in cell culture medium and total RNA was extracted from the cell pellets using the TRIzol reagent as described by the manufacturer (Invitrogen, Carlsbad, CA). For MDA-MB-435 xenografts in athymic nude mice, tumors were removed at necropsy, immediately placed in RNALater™ (Ambion, Austin, TX) and stored at -80˚C as previously described (12). Frozen xenografts from mice were placed in '1x1' plastic bags, pulverized on dry ice, transferred to 35 ml conical Oakridge tubes (Nalgene, Rochester, NY), and weighed. Frozen tissues were homogenized in TRIzol using a polytron homogenizer (Brinkmann Instruments, Inc. Westbury, NY) and total RNA isolated using the TRIzol reagent. For the human tumors, frozen tissue was placed in a '1x1' plastic bag on dry ice, pulverized, and lysis buffer added (Qiagen RNeasy kit; Qiagen Inc., Valencia, CA). Each sample was then homogenized with a 1 ml syringe and 18 gauge needle, added to the Qiagen spin column, processed as described by the manufacturer, and RNA eluted with dH_2O. None of the RNAs was amplified or pooled.

RNA concentrations were determined by comparing the optical density ratios (OD_{260}/OD_{280}) obtained spectrophotometrically using a Beckman DU640 Spectrophotometer (Beckman, Fullerton, CA). RNA quality was assessed using an Agilent 2100 Bioanalyzer and RNA 6000 LabChip kits (Agilent Technologies, New Castle, DE), which allows for visual examination of both the 18S and 28S rRNA bands as a measure of RNA integrity. We used high quality RNA as assessed by standard measures (12).

NamedGenes GeneFilters (ResGen, part of the Invitrogen Corporation, Inc., Huntsville, AL) that contain 4,132 known cDNAs and 192 controls including total genomic DNAs (tgDNA) on each filter were used. Probes were generated as previously described (14). Briefly, total RNA (500 ng) from experimental samples was reverse transcribed and simultaneously radioactively labeled by incorporation of [α-^{33}P]ATP and [α-^{33}P]CTP. This method radiolabels both the sense and antisense probe strands. Probes were purified and hybridized to a GeneFilter, and incubated for 12-18 h at 42˚C in a roller oven (Robbin Scientific, Sunnyvale, CA). Each hybridized GeneFilter was washed twice in 2X SSC, 1% SDS at 50˚C for 20 min and once at 55˚C in 0.5X SSC, 1% SDS for 15 min. Hybridization signals were detected by phosphorimage analysis using a Molecular Dynamics Storm PhosphorImager (Molecular Dynamics, Sunnyvale, CA).

Microarray data preprocessing. Pathways™ 4.0 software algorithms (Research Genetics, Inc.) were used to acquire data from microarray images. Briefly, this software geometrically quantifies the intensities of both the spot and local background for each gene. Local background correction is estimated by subtracting local signals from areas devoid of target from the raw intensity value of each target cDNA, and a value of one is added to all non-negative values to conserve the relative intensities with low expression values. Negative values resulting from background subtraction were adjusted to one. Background-corrected data were then normalized to account for differences in probe specific activity, hybridization, and other variables among replicates. The global mean method was used to normalize the data from each array.

A signal bleeding effect from neighboring cDNA spots, where signals from adjacent spots bleed into each other, is a major confounding factor with this microarray technology. To determine if a spot on the filter was affected by signal bleeding, we used an in-house algorithm (programmed in MatLab version R13SP1; Mathworks, Natick, MA; unpublished data). This algorithm calculates the difference between the respective local background for a gene and global background from the filter, expressed as a percentage of the raw intensity value for that gene. Values above a predetermined threshold indicate that the signal from neighboring spots bled into the spot of interest. The digitized images for all spots flagged by the algorithm were subjected to visual inspection to confirm any signal bleeding. Genes with signals determined visually and/or mathematically to be confounded by a bleed effect were excluded from further analysis.

We used several criteria to identify and exclude likely non-informative genes and construct a reduced dimensional data set for analysis. The goal of these preprocessing steps was to obtain a series of robust expression values for genes determined to be present in all three groups to be compared in the study (3 ER+ human breast cancer cell lines; 13 breast tumors; 38 MDA-MB-435 xenografts). First, we excluded genes that have expression values consistently in the undetectable range in all microarrays or that have signals compromised by signal bleed as defined above. If a gene was found to be free of bleeding effects in at least 70% of arrays, data for this gene were retained for further study. Genes in the undetectable range were eliminated if their normalized expression levels were <0.1 in all experimental groups. We did not attempt to estimate and replace missing values. Application of these criteria across all microarrays from the cell lines, breast tumors, and MDA-MB-435 xenografts resulted in a list of 428 robust gene signals for further analysis.

Data analysis: comparison of high dimensional data structures. To estimate independently the data structures, we conducted separate PCA on the robust gene expression data set (n=428 genes) for each of the three groups and determined the essential dimensionality (M) for samples within the same group. PCA was performed using the covariance matrices for standardized gene expression levels. M is defined as the number of principal components (PCs) needed to account for the variation in the original data. Jolliffe proposed several strategies to determine M (11); we applied the most commonly used rule and selected those PCs that represent the smallest value of M that captures a high cumulative percentage of the total variance (≥80%).

Once the M PCs were identified for a group, we calculated the Pearson's correlation coefficient for each gene with each PC and selected those genes with an absolute correlation coefficient r≥0.800 with at least one of the top M PCs (top genes). While this approach is broadly comparable to the method proposed by Jolliffe (11), we ranked the PCs such that PC1 captured the highest proportion of data variance, PC1 + PC2 captured the next highest proportion and we continued until PC1 + PC2...PC$_i$ captured ≥80% of the total variance. Thus, we placed more weight on the top PCs, whereas Jolliffe's method attributed equal importance to each of the M PCs. Our approach appears reasonable, since genes tend to have larger correlation coefficients with higher ranked compared with lower ranked PCs.

Since we explored independently each group, the PCs and the genes that best define these PCs reflect only the structure of the data for that group. In this manner, we can compare the relative importance of each gene expression value across data structures. Thus, having selected the top genes from each of the three groups, we compared the respective M PC-derived gene lists among groups and created a 'common genes' list. For example, if *gene-1* was one of the top genes for both breast tumor and cell line samples, we considered *gene-1* as a common gene between these two groups.

Results

Cell line and tumor data structures share similar essential dimensionality. For this study, an unsupervised probabilistic approach applied to each experimental group should have the greatest potential to generate relatively unbiased, independent representations of data structures. Since we do not predetermine the number of PCs, only the percentage of cumulative variation, and the M PCs are independently identified within each group, the M PCs obtained should provide reasonable representations by which to compare data structures. While we might expect similar data structures to be defined by approximately similar numbers of M PCs, this is an inadequate single measure because the genes most closely correlated with each PC may be different. Conversely, it is possible that a different number of PCs may be required to satisfy M in each experimental group but the genes correlated with the respective M PCs may be very similar.

To address these issues, we compared the number of M PCs, ranked these by their relative ability to capture data variation, and then assessed the correlation of each gene with each ranked PC. Data sets that exhibit similarities may be defined by similar numbers of M PCs. More importantly, data structures with substantial similarities will have the same genes highly correlated with similarly ranked PCs; for example, *gene-1* is highly correlated with PC1 in one group and also is highly correlated with PC1, PC2, or PC3 in another experimental group. The higher proportions of genes

Table II. Principal component analysis and essential dimensionality.

PCA	Cell lines (%)	Tumors (%)	Cell lines/tumors (combined) (%)
M PCs	n=5	n=6	n=10
PC1	31.8	35.8	28.8
PC1 + PC2	51.0	48.8	42.9
PC1…PC3	65.7	58.2	50.5
PC1…PC4	74.6	60.7	57.0
PC1…PC5	**80.9**	74.5	62.6
PC1…PC6	85.7	**81.4**	67.5
PC1…PC7	89.9	87.1	71.8
PC1…PC8	93.0	91.3	75.9
PC1…PC9	95.8	94.2	79.1
PC1…PC10	98.1	96.5	**81.8**
r≥0.800 (M PC)	n=103 genes	n=106 genes	n=65 genes
Common genes	36 genes are common to the 103-genes (cell lines) and 106-genes (tumors)		31 of the 36 common genes correlate with PC1 of the combined group[a]

M PCs is the number of PCs required to capture ≥80% of the cumulative variances in the data set (essential dimensionality). Percentages are the cumulative variances captured by the sum of the M PCs as indicated. The final row shows the number of genes in each group that have a correlation coefficient r≥0.800 with at least one of the M PCs in that group. For example, there are 103 genes correlated either with PC1, PC2, PC3, PC4 or PC5 in the breast cancer cell line data set. [a]For 22 genes r≥0.800; for a further 9 genes r≥0.750.

that are highly correlated with top ranked PCs in both groups, the more similar are the data structures being compared.

In this data set there are over 420 possible orthogonal PCs that can be explored as projections of the high dimensional data. However, we would expect most of the data variation to be captured by a much smaller number of M PCs. Using our approach, we found that only six PCs are required to define the breast tumor data structure by our criterion of ≥80% cumulative variance (cumulative variance = 80.9%; Table II). Similarly, only five PCs are required to describe the breast cancer cell line data structure (cumulative variance = 81.4%; Table II).

The top principal components of cell line and tumor data share notable similarities. To compare the PCs, we then calculated the correlation coefficient of each gene with each of the M PCs and obtained two gene lists, one for each experimental group. Thirty-six genes are important in describing independently the data structures for both the tumor and cell line groups (Table III). Surprisingly, all 36 common genes were correlated with the top ranked PC (PC1) from the tumor data set. Of the genes from the ER^+ cell lines data set, 21 genes also correlated with its PC1. Thus, there are striking similarities between the top PC in both data sets; each of which capture almost one-third of the variation in their respective data sets (Table II). Of the remaining 15 genes, 14 genes are correlated with PC2; only one gene is correlated with PC3 in the ER^+ cell lines data. The sign of the correlation is less informative than the absolute value of the coefficient; since we would not expect the PCs to be identical, the direction of each gene's correlation with a PC may vary in each data set and its absolute value reflects the true significance. Nonetheless, 61% of the genes (22/36) show the same directional correlation. Twenty-one of these genes correlated with PC1, strongly suggesting substantial similarities in the top PC. Taken together, these data provide evidence of notable similarities between the human breast cancer cell line and breast tumor transcriptome data structures.

To further support these observations, we combined the cell line and tumor data sets and performed PCA on the combined group. Since the tumors are more heterogeneous than the cell lines, we would expect the combined data set to require a higher number of M PCs and that fewer of the previously identified common genes will be highly correlated with these M PCs. Consistent with the general similarities, Table II shows that only 10 PCs are required (cumulative variance = 81.8%) to define the structure of the combined data set. We then calculated the correlation coefficients for the previously identified 36 common genes with the top PCs derived from the combined group. Thirty-five genes could be evaluated since one gene was not correlated with the top PCs. Twenty-two of the common genes met the initial criterion of $r \geq 0.800$ and a further 9 genes had correlations of $r \geq 0.750$ (Table III). All 31 of these genes were correlated with PC1. The remaining 4 genes were correlated with PC2 (n=1) or PC3 (n=3) but their coefficients were much lower. Thus, most of the common genes important in separate group analysis also are important in combined group analysis.

Since PCA can be used to perform multidimensional scaling for visualization (12,15), we used the top two PCs to visualize the combined data group. Fig. 1 shows that the cell line and breast tumor samples do not form distinct separable clusters in 2-dimensional PC space. These projections are visually consistent with the PCA analysis described above.

We also performed similar independent M PC analyses using data from 38 MDA-MB-435 xenografts growing in the mammary fat pad regions of athymic nude mice (data not shown). Capturing the essential dimensionality of the data structure required 16 PCs and no genes met the criteria for commonality between these xenografts and the breast tumors. Only four genes were found to be in common with the three ER^+ breast cancer cell lines: S100A11 (S100 calcium binding protein A11), PTPN7 (protein tyrosine phosphatase, non-receptor type 7), MR1 (major histocompatibility complex, class I-related), and DCI (dodecenoyl-Coenzyme A delta isomerase; 3,2 trans-enoyl-Coenzyme A isomerase). The notable lack of similarity with the breast cancer cell lines and tumors is consistent with the putative non-breast cancer origin (16), although the ER^- status of this cell line and the predominantly ER^+ status of the breast tumors and breast cancer cell lines also may contribute to the lack of similarity in MDA-MB-435 xenograft data structure with the breast tumors and data sets of other cell lines.

While our approach was not designed to select genes for their functional relevance or differential association among specific breast cancer outcomes/phenotypes, we might expect some of these genes to represent functions implicated in other breast cancer studies. We used the six main gene ontology functional categories as defined in the GO database (http://www.geneontology.org) and applied by Pawitan *et al* (17), who compared two gene lists implicated in predicting breast cancer prognosis. This appears to be a reasonable comparison as our data set included both lymph node positive and negative cases (Table I); lymph node involvement is one of the strongest independent predictors of a poor prognosis (18,19). Since there are only three common genes between the 64-gene (Pawitan) and 70-gene (van't Veer) gene lists, despite the similarities between these two studies, it was not surprising that we did not find any of those genes in common with our 36 genes. However, we found 11/36 genes in 5 of the 6 functional categories (Table IV). Thus, 31% of the genes are represented in the 6 functional categories, compared with 37% of the van't Veer *et al* genes (20) and 45% of the Pawitan *et al* genes (21).

Discussion

Limitations in the ability of individual experimental models to reflect fully the complexity of their corresponding human cancer are widely acknowledged. For example, cellular signaling in rodent cells may not be similar to that in human cells. Human cells require notably more changes in genetic, epigenetic, or gene expression events for malignant transformation (22-24); the same may be true for post-transformation events that drive malignant progression. While established human breast cancer cell lines exhibit many phenotypic characteristics of the human disease, the ability to use these models to discover meaningful molecular insights into breast cancer biology also remains controversial (8). Thus, the primary goal of this study was to compare the transcriptome structures, as derived from gene expression microarray data,

Table III. Common genes correlated with the top M PCs in the breast tumors and cell lines.

Gene	Gene name	Cells	r	Tumors	r	Comb	R
METAP2	Methionyl aminopeptidase 2	PC2	0.964	PC1	-0.857	-0.394	PC3
A2M	Alpha-2-macroglobulin	PC1	-0.885	PC1	-0.835	-0.804	PC1
IGFBP6	Insulin-like growth factor binding protein 6	PC2	0.956	PC1	-0.900	0.844	PC1
KRT13	Keratin 13	PC2	0.869	PC1	-0.844	-0.822	PC1
DRAP1	DR1-associated protein 1 (negative cofactor 2 alpha)	PC2	0.979	PC1	-0.908	0.754	PC1
GPC1	Glypican 1	PC1	-0.805	PC1	-0.957	-0.901	PC1
PCOLN3	Procollagen (type III) N-endopeptidase	PC1	-0.825	PC1	-0.805	-0.786	PC1
ATP5J	ATP synthase, H$^+$ transporting, mitochondrial F0 complex, subunit F6	PC2	0.960	PC1	-0.874	0.778	PC1
MRPL49	Mitochondrial ribosomal protein L49	PC1	-0.875	PC1	-0.927	-0.903	PC1
RELA	NFκB (p65)	PC2	0.969	PC1	-0.861	-0.758	PC1
PTHR1	Parathyroid hormone receptor 1	PC1	-0.886	PC1	-0.857	-0.793	PC1
FST	Follistatin	PC2	0.930	PC1	-0.881	-0.349	PC3
POLA	Polymerase (DNA directed), alpha	PC1	-0.854	PC1	-0.858	-0.826	PC1
CREBL1	cAMP responsive element binding protein-like 1	PC1	-0.937	PC1	-0.862	-0.873	PC1
GOLGA2	Golgi autoantigen, golgin subfamily a, 2	PC2	0.809	PC1	-0.816	0.562	PC3
SF3A1	Splicing factor 3a, subunit 1, 120 kDa	PC2	0.962	PC1	-0.841	0.494	PC2
USP4	Ubiquitin specific protease 4 (proto-oncogene)	PC1	-0.889	PC1	-0.912	-0.898	PC1
CR2	Complement component (3d/Epstein-Barr virus) receptor 2	PC1	-0.835	PC1	-0.818	-	-
NR1D1	Nuclear receptor subfamily 1, group D, member 1	PC1	-0.885	PC1	-0.874	-0.884	PC1
ODC1	Ornithine decarboxylase 1	PC1	-0.870	PC1	-0.945	-0.901	PC1
ORM2	Orosomucoid 2	PC2	0.962	PC1	-0.972	0.876	PC1
AMFR	Autocrine motility factor receptor	PC1	-0.824	PC1	-0.883	-0.887	PC1
RYR1	Ryanodine receptor 1	PC2	0.981	PC1	-0.927	-0.772	PC1
PPM1F	Protein phosphatase 1F	PC1	-0.820	PC1	-0.845	-0.843	PC1
KCNN4	Potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4	PC1	-0.802	PC1	-0.929	-0.886	PC1
NT5E	5' nucleotidase (CD73)	PC2	0.912	PC1	-0.819	-0.786	PC1
ITGB2	Integrin beta 2	PC1	-0.908	PC1	-0.885	-0.876	PC1
ABCC1	ATP-binding cassette, subfamily C (CFTR/MRP), member 1	PC2	0.953	PC1	-0.923	-0.827	PC1
PXN	Paxillin	PC1	-0.889	PC1	-0.951	-0.929	PC1
STAM	Signal transducing adaptor molecule (SH3 domain and ITAM motif) 1	PC2	0.899	PC1	-0.882	-0.817	PC1
COX6B1	Cytochrome c oxidase subunit VIb	PC3	-0.813	PC1	-0.980	-0.895	PC1
ACTR1A	Actin-related protein 1 homolog A	PC1	-0.837	PC1	-0.906	-0.882	PC1
LOC56311	Ankyrin repeat domain 7	PC1	-0.911	PC1	-0.854	-0.796	PC1
KIAA1641	Chronic lymphocytic leukemia-associated antigen KW-1	PC1	-0.845	PC1	-0.907	-0.797	PC1
CSTA	Cystatin A (stefin A)	PC1	-0.892	PC1	-0.928	-0.910	PC1
B7	B7 protein	PC1	-0.845	PC1	-0.850	-0.804	PC1

For comparison of the cell lines and tumors, each gene selected must exhibit a correlation coefficient of r≥0.800 with one of the top M PCs. For example, a gene in common between breast tumors and breast cancer cell lines must be correlated (r≥0.800) with PC1, PC2, PC3, PC4 or PC5; there are only 5 M PCs in the breast cancer cell line group (see Table II); there are 36 genes in common by these criteria. The gene CR2 was not associated with the top M PCs in the combined group. Gene, gene symbol as designated by the human gene ontology (HUGO) gene nomenclature committee. Comb, data from the combined cell line (MCF-7, T47D, ZR-75-1) and tumor data set. The four genes in the combined data set where r<0.75 are indicated.

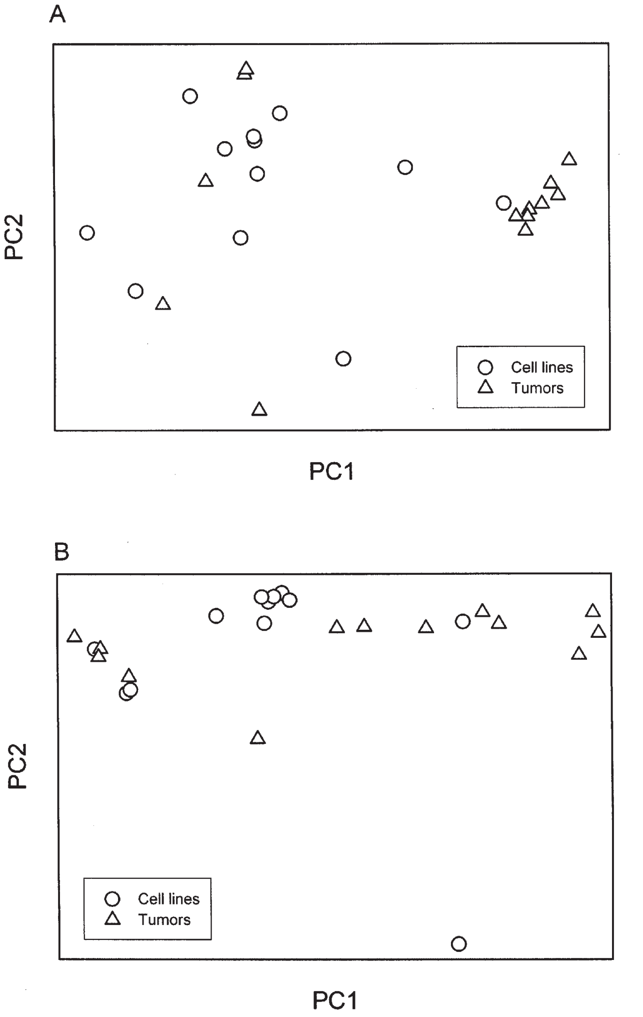

Figure 1. Multidimensional scaling of cell line and tumor data. (A), 428 dimensional data set; (B), 36 dimensional data set. Δ, breast tumors; O, cell lines.

Table IV. Gene functions among the 36 common genes.

Biological function	Genes
DNA replication/transcription	CREBL1, NR1D1, POLA, RELA, RYR1
Apoptosis	ITGB2, PPM1F, RELA, RYR1
Cell cycle/proliferation/growth	IGFBP6, RYR1
Cell adhesion/motility	AMFR, ITGB2, PXN, RYR1
Signal transduction	AMFR, CREBL1, IGFBP6, ITGB2, PTHR1, PXN, RELA, RYR1, STAM

The GO database was used to annotate the gene functions (http://www.geneontology.org). The GO categories are based on the six used by Pawitan *et al* (17) to compare their breast cancer predictive gene list with that of van't Veer *et al* (20). We found no genes in the 'angiogenesis' category; Pawitan *et al* reported only one gene from their 64-gene data set and a different gene from the 70-gene van't Veer *et al* data set in this category (17).

of predominately ER$^+$ breast tumor specimens from patients and the three most widely used ER$^+$ human breast cancer cell lines (MCF-7, T47D, ZR-75-1).

Breast tumor specimens can include multiple different cell types such as epithelial, myoepithelial, fibroblastic, myofibroblastic, and reticuloendothelial (25), whereas cell lines are, in comparison, biologically more homogeneous. Thus, the goal of comparing cell lines and tumor specimens, using direct comparisons of gene expression levels, is potentially confounded by tissue heterogeneity. In breast tumors, a gene's signal will reflect the sum of values from all cell types included in the specimen. Earlier microarray studies did not account for this heterogeneity and this may partly explain the greater similarity reported between normal breast and breast cancer specimens than between the breast cancer specimens and human breast cancer cell lines (6). Furthermore, earlier studies used unsupervised hierarchical clustering methods to solve the high dimensional data structures and identify putative relationships among samples. Since these hierarchies can be built using different distance measures and the data points linked by different measures (26-28), different clustering methods may provide different solutions to the same data sets (29,30). With no goodness-of-fit for the data solutions (29) or comparisons with other methods that may provide more accurate or more complete solutions, the inability of breast cancer cell lines to cluster together or to cluster with breast cancers may reflect the limitations inherent in the analytical approaches applied. The lack of consideration of specimen heterogeneity also may have confounded the analysis.

Rather than apply heuristic rules to deduce similarities or differences based on broad phenotypic characteristics or other observations, we applied a relatively unbiased probabilistic approach to compare transcriptomes. Unlike most prior microarray studies that focus upon finding differential gene expression patterns among groups, we were most interested in those genes that are commonly important in defining data structure. While we would expect differences in the absolute levels or patterns of expression of some genes, our main goal was to explore the similarities in overall data structures. Differences in absolute gene expression values could lead to the appearance of differential gene expression values that may more closely reflect the cellular rather than molecular differences between relatively homogeneous cell lines and heterogeneous tumors.

The probabilistic approach we used compares the M PC projections in the data sets and those genes that best define these respective PCs. Thus, the method should capture, in a largely unbiased manner, those PCs and genes that best define the structure of each high dimensional data set - at least as defined by its total variation. Our data show that the three most widely used ER$^+$ human breast cancer cell lines, even when growing *in vitro*, exhibit marked similarities to a panel of ER$^+$ breast tumor specimens. These molecular observations on the primary structure of the breast tumor and cell line transcriptomes appear consistent with the widely reported biological similarities between these cell lines, their variants, and the human disease (2-4).

The genes identified in these tumors and models reflect the specimens and microarray technology used; similar data collected from other breast tumors, cell lines, or microarray platforms may or may not find the same genes to define the M PCs of those data sets. However, we would anticipate that such studies may find genes that exhibit similar statistical properties, or perhaps broadly similar molecular functions, to be associated with the top PCs. Since we identified genes that best define the PCs from a small but robust subset of expression measures, their selection reflects a probabilistic assessment only of their contribution to global data structure. Thus, there is no compelling biological rationale why these specific genes must reflect key biological processes in breast tumors. The use of PCA for gene selection in mechanistic studies is potentially flawed for several reasons, some of which are discussed elsewhere (31). Nonetheless, it is intuitively reasonable to expect some genes closely associated with data structure to broadly reflect key molecular processes and/or include genes already implicated in breast cancer.

Several of the genes or gene functions represented in the 36 common genes identified herein have been directly or indirectly implicated in affecting key breast cancer phenotypes. For example, we found 11 genes in 5 of the 6 gene function categories implicated in separating good prognosis from poor prognosis breast cancers (17). While our study would not be expected to find the same genes as these two previous studies - we did not look for such discriminant genes nor did we use similar microarray platforms - the data in Table IV suggest that the 36 common genes and/or the functional categories they represent are important in both human breast cancer and human breast cancer cell lines. Examples of specific genes from the 36 common gene list include RELA (NFκB p65), ornithine decarboxylase-1 (ODC1), paxillin (PXN), and insulin-like growth factor (IGF) binding protein-6 (IGFBP-6). RELA is implicated in estrogen independence (32,33) and acquired antiestrogen resistance in cell culture models (15,34,35), and is readily detected by immunohistochemistry in breast tumors (36). The polyamine ODC1 is estrogen regulated (37,38), is a target for drug development (38,39), and is a potential breast cancer biomarker (40). The focal adhesion protein PXN is regulated by heregulin, a key effector of breast cancer cell growth (16). PXN expression also is regulated by activation of the IGF-type 1 receptor (41). This receptor is activated by IGF-II, a major mitogen for breast cancer cells (42); IGFBP6 has a notably high affinity for binding IGF-II, inhibits its activity (43), and also is a candidate breast cancer biomarker (21).

The data we present here suggest that well-established ER$^+$ human breast cancer cell lines and breast tumors share global similarities in the structures of their respective transcriptomes. The strong correlations of similar genes with the top PC projections in each data set clearly imply that MCF-7, T47D, and ZR-75-1 cells are good models in which to identify molecular events that also are important in some ER$^+$ human breast cancers.

Acknowledgments

This work was supported in part by Public Health Service award R01-CA096483 (R. Clarke), and a Clinical Translational Award (W81XWH-04-1-0570) from the United States Army Medical Research and Materiel Command Breast Cancer Research Program (R.C.), BC030280 (R.C.). Technical services also were provided by the Histopathology Shared Resource funded through Public Health Service award P30-CA51008-14 (Lombardi Comprehensive Cancer Center Support Grant). The authors wish to thanks Dr Rebecca Riggins for critical reading of the manuscript.

References

1. Ceriani RL, Peterson JA, Blank EW, Chan CM and Cailleau R: Development and characterization of breast carcinoma cell lines as *in vivo* models for breast cancer diagnosis and therapy. In Vitro Cell Dev Biol 28A: 397-402, 1992.
2. Clarke R: Human breast cancer cell line xenografts as models of breast cancer. The immunobiologies of recipient mice and the characteristics of several tumorigenic cell lines. Breast Cancer Res Treat 39: 69-86, 1996.
3. Clarke R, Leonessa F, Welch JN and Skaar TC: Cellular and molecular pharmacology of antiestrogen action and resistance. Pharmacol Rev 53: 25-71, 2001.
4. Lacroix M and Leclercq G: Relevance of breast cancer cell lines as models for breast tumours: an update. Breast Cancer Res Treat 83: 249-289, 2004.
5. Price JE, Polyzos A, Zhang RD and Daniels LM: Tumorigenicity and metastasis of human breast carcinoma cell lines in nude mice. Cancer Res 50: 717-721, 1990.
6. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de RM, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D and Brown PO: Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24: 227-235, 2000.
7. Van't Veer LJ and Weigelt B: Road map to metastasis. Nat Med 9: 999-1000, 2003.
8. Burdall SE, Hanby AM, Lansdown MR and Speirs V: Breast cancer cell lines: friend or foe? Breast Cancer Res 5: 89-95, 2003.
9. Ross DT and Perou CM: A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. Dis Markers 17: 99-109, 2001.
10. Hotelling H: Analysis of a complex of statistical variables into principal components. J Educ Psychol 24: 417-441 and 498-520, 1933.
11. Jolliffe IT: Principal Component Analysis. 2nd edition. Springer-Verlag, New York, NY, 2002.
12. Ellis M, Davis N, Coop A, Liu M, Schumaker L, Lee RY, Srikanchana R, Russell CG, Singh B, Miller WR, Stearns V, Pennanen M, Tsangaris T, Gallagher A, Liu A, Zwart A, Hayes DF, Lippman ME, Wang Y and Clarke R: Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses. Clin Cancer Res 8: 1155-1166, 2002.
13. James MR, Skaar TC, Lee RY, MacPherson A, Zwiebel JA, Ahluwalia BS, Ampy F and Clarke R: Constitutive expression of the steroid sulfatase gene supports the growth of MCF-7 human breast cancer cells *in vitro* and *in vivo*. Endocrinology 142: 1497-1505, 2001.
14. Sgroi DC, Teng S, Robinson G, Le Vangie R, Hudson JR and Elkahloun AG: *In vivo* gene expression profile analysis of human breast cancer progression. Cancer Res 59: 5656-5661, 1999.
15. Gu Z, Lee RY, Skaar TC, Bouker KB, Welch JN, Lu J, Liu A, Zhu Y, Davis N, Leonessa F, Brunner N, Wang Y and Clarke R: Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-kappaB and cyclic AMP response element binding with acquired resistance to faslodex (ICI 182,780). Cancer Res 62: 3428-3437, 2002.
16. Rae JM, Ramus SJ, Waltham M, Armes JE, Campbell IG, Clarke R, Barndt RJ, Johnson MD and Thompson EW: Common origins of MDA-MB-435 cells from various sources with those shown to have melonoma properties. Clin Exp Metastasis 21: 543-552, 2004.
17. Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedren S and Bergh J: Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 7: R953-R964, 2005.

18. Carter CL, Allen C and Henson DE: Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. Cancer 63: 181-187, 1989.
19. Fisher B, Bauer M, Wickerham DL, Redmond CK, Fisher ER, Cruz AB, Foster R, Gardner B, Lerner H and Margolese R: Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update. Cancer 52: 1551-1557, 1983.
20. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der KK, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R and Friend SH: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530-536, 2002.
21. Kaulsay KK, Ng EH, Ji CY, Ho GH, Aw TC and Lee KO: Serum IGF-binding protein-6 and prostate specific antigen in breast cancer. Eur J Endocrinol 140: 164-168, 1999.
22. Rangarajan A, Hong SJ, Gifford A and Weinberg RA: Species- and cell type-specific requirements for cellular transformation. Cancer Cell 6: 171-183, 2004.
23. Rangarajan A and Weinberg RA: Opinion: comparative biology of mouse versus human cells: modelling human cancer in mice. Nat Rev Cancer 3: 952-959, 2003.
24. Hahn WC and Weinberg RA: Modelling the molecular circuitry of cancer. Nat Rev Cancer 2: 331-341, 2002.
25. Clarke R, Dickson RB and Lippman ME: Hormonal aspects of breast cancer: growth factors, drugs and stromal interactions. Crit Rev Oncol Hematol 12: 1-23, 1992.
26. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R and Staudt LM: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511, 2000.
27. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO and Weinstein JN: A gene expression database for the molecular pharmacology of cancer. Nat Genet 24: 236-244, 2000.
28. Thykjaer T, Workman C, Kruhoffer M, Demtroder K, Wolf H andersen LD, Frederiksen CM, Knudsen S and Orntoft TF: Identification of gene expression patterns in superficial and invasive human bladder cancer. Cancer Res 61: 2492-2499, 2001.
29. Satagopan JM and Panageas KS: A statistical perspective on gene expression data analysis. Stat Med 22: 481-499, 2003.
30. Hinneburg A and Keim DA: Optimal grid-clustering: towards breaking the curse of dimensionality in high-dimensional clustering. In: Proceedings of the 25th Conference on Very Large Databases. Atkinson MP, Orlowska ME, Valduriez P, Zdonik SB and Brodie ML (eds.) Morgan Kaufman, San Francisco, pp506-517, 1999.
31. Wittes J and Friedman HP: Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. J Natl Cancer Inst 91: 400-401, 1999.
32. Nakshatri H, Bhat-Nakshatri P, Martin DA, Goulet RJ and Sledge GW: Constitutive activation of NF-kappaB during progression of breast cancer to hormone-independent growth. Mol Cell Biol 17: 3629-3639, 1997.
33. Pratt MAC, Bishop TE, White D, Yasvinski G, Menard M, Niu MY and Clarke R: Estrogen withdrawal-induced NF-kappaB activity and bcl-3 expression in breast cancer cells: roles in growth and hormone independence. Mol Cell Biol 23: 6887-6900, 2003.
34. Riggins R, Zwart A, Nehra N, Agarwal P and Clarke R: The NFκB inhibitor parthenolide restores ICI 182,780 (Faslodex; Fulvestrant)-induced apoptosis in antiestrogen resistant breast cancer cells. Mol Cancer Ther 4: 33-41, 2005.
35. Osipo C, Gajdos C, Liu H, Chen B and Jordan VC: Paradoxical action of fulvestrant in estradiol-induced regression of tamoxifen-stimulated breast cancer. J Natl Cancer Inst 95: 1597-1608, 2003.
36. Zhu Y, Singh B, Hewitt S, Liu A, Gomez B, Wang A and Clarke R: Expression patterns among interferon regulatory factor-1, human X-box binding protein-1, nuclear factor kappa B, nucleophosmin, estrogen receptor alpha and progesterone receptor proteins in breast cancer tissue microarrays. Int J Oncol 28: 67-76, 2006.
37. Qin C, Samudio I, Ngwenya S and Safe S: Estrogen-dependent regulation of ornithine decarboxylase in breast cancer cells through activation of nongenomic cAMP-dependent pathways. Mol Carcinog 40: 160-170, 2004.
38. Manni A: Polyamine involvement in breast cancer phenotype. In Vivo 16: 493-500, 2002.
39. Manni A, Washington S, Mauger D, Hackett DA and Verderame MF: Cellular mechanisms mediating the anti-invasive properties of the ornithine decarboxylase inhibitor alpha-difluoromethylornithine (DFMO) in human breast cancer cells. Clin Exp Metastasis 21: 461-467, 2004.
40. Mimori K, Mori M, Shiraishi T, Tanaka S, Haraguchi M, Ueo H, Shirasaka C and Akiyoshi T: Expression of ornithine decarboxylase mRNA and c-myc mRNA in breast tumours. Int J Oncol 12: 597-601, 1998.
41. Guvakova MA and Surmacz E: The activated insulin-like growth factor I receptor induces depolarization in breast epithelial cells characterized by actin filament disassembly and tyrosine dephosphorylation of FAK, Cas, and paxillin. Exp Cell Res 251: 244-255, 1999.
42. Sachdev D and Yee D: The IGF system and breast cancer. Endocr Relat Cancer 8: 197-209, 2001.
43. Bach LA: IGFBP-6 five years on; not so ‘forgotten’? Growth Horm IGF Res 15: 185-192, 2005.